

Coalescence theory as a tool for population genetics

```
co•a•lesce | kōəˈles|
verb [intrans.]
come together and form one mass or whole : the puddles had
coalesced into shallow streams | the separate details coalesce to
form a single body of scientific thought.
```

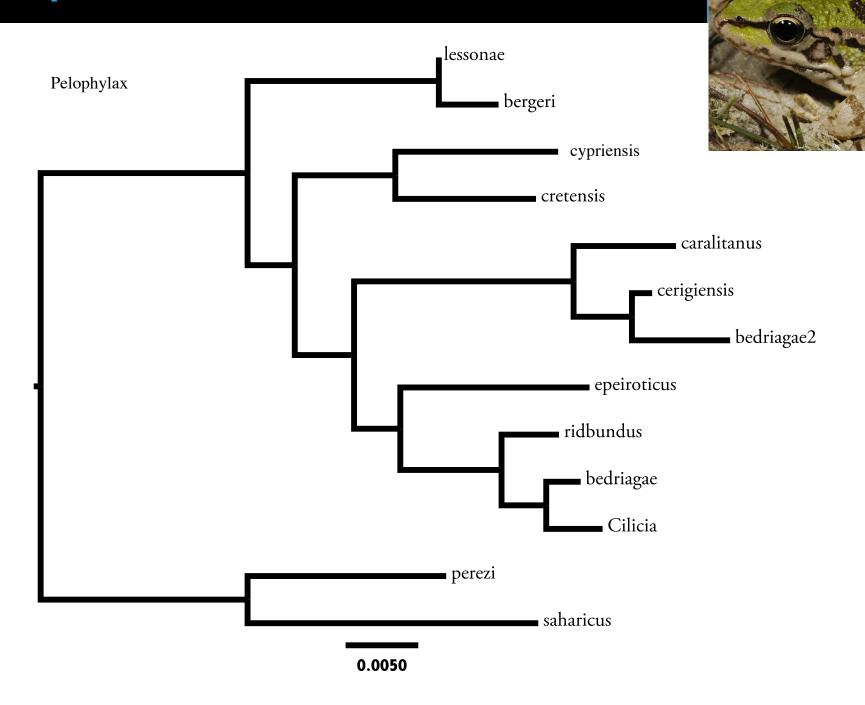
• [trans.] combine (elements) in a mass or whole : to help coalesce the community, they established an office.

DERIVATIVES

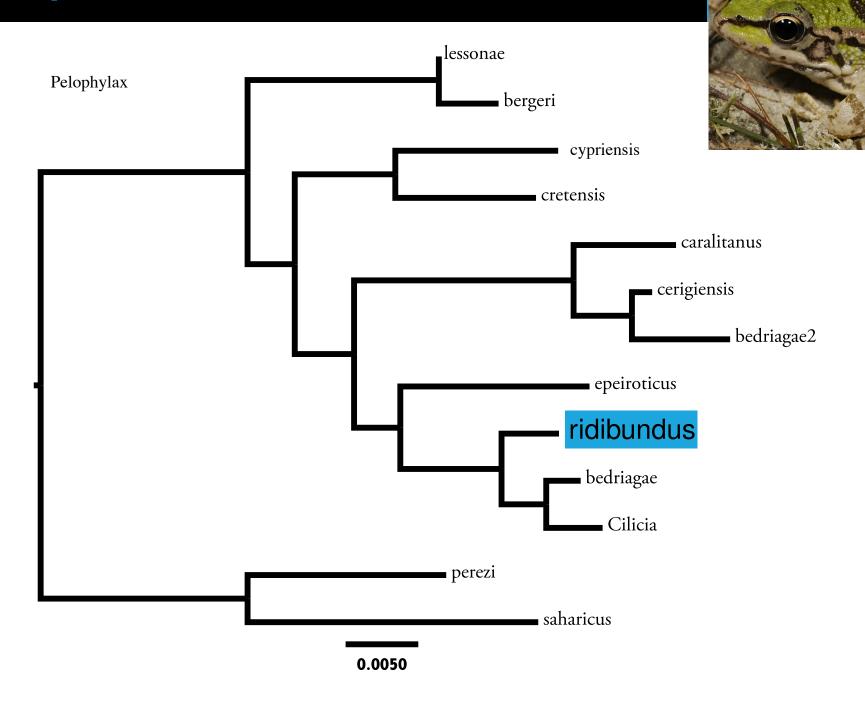
```
co•a•les•cence |-'lesəns| noun
co•a•les•cent |-'lesənt| adjective
```

ORIGIN mid 16th cent. (in the sense [bring together, unite]): from Latin *coalescere*, from *co-* (from *cum 'with'*) + *alescere 'grow up'* (from *alere 'nourish'*).

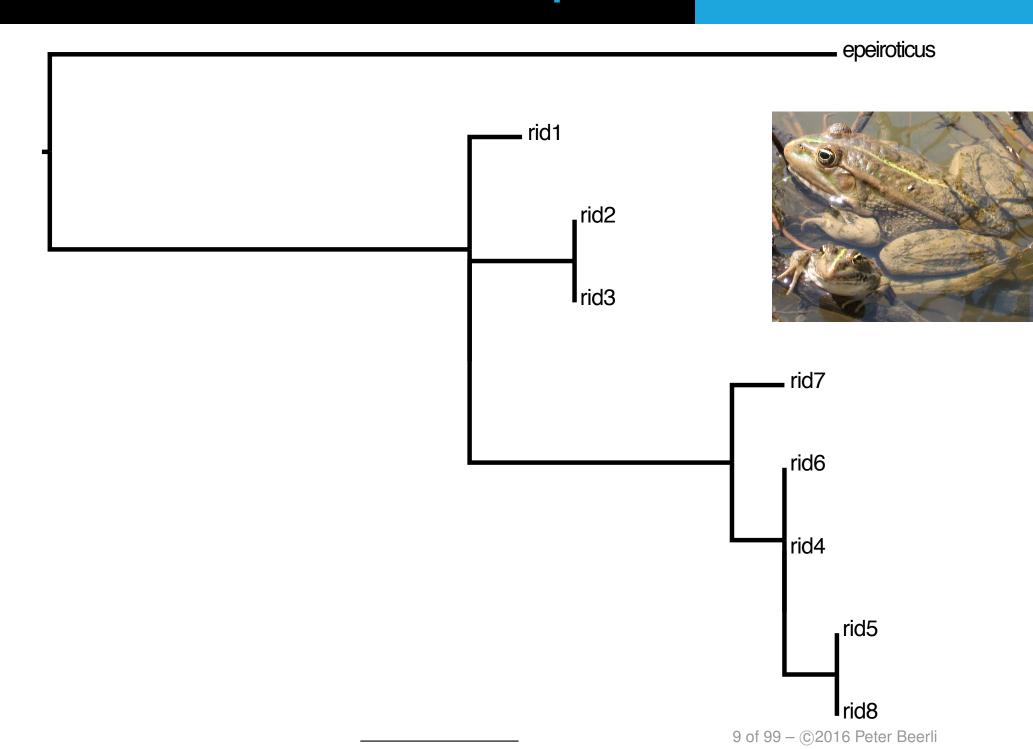
Species trees



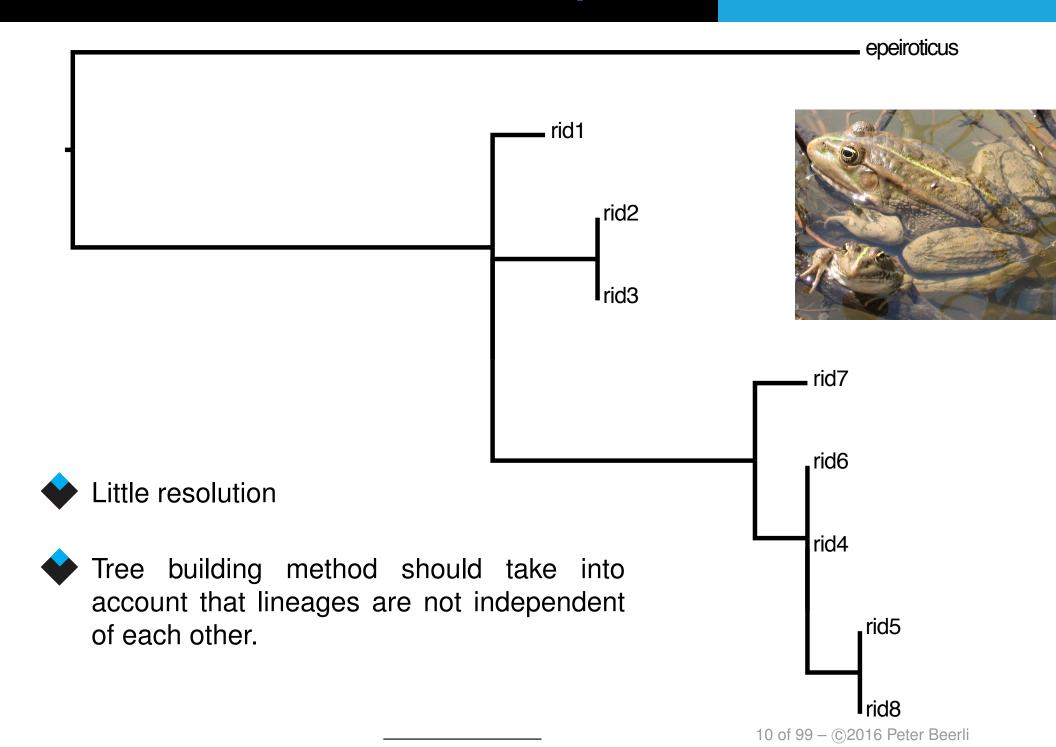
Species trees



Tree of individuals of same species



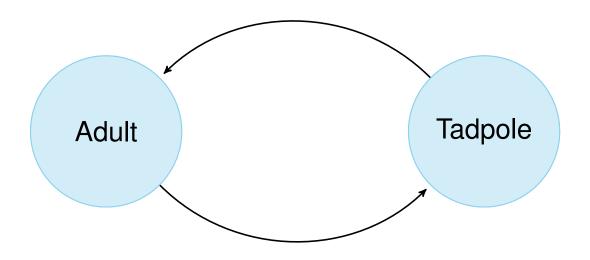
Tree of individuals of same species



Interaction among individuals



Interaction among individuals



Wright-Fisher population model

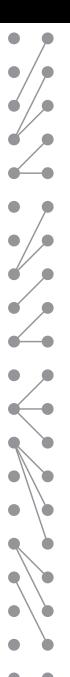
- All individuals live one generation and get replaced by their offspring
- All have same chance to reproduce, all are equally fit
- The number of individuals in the population is constant

•

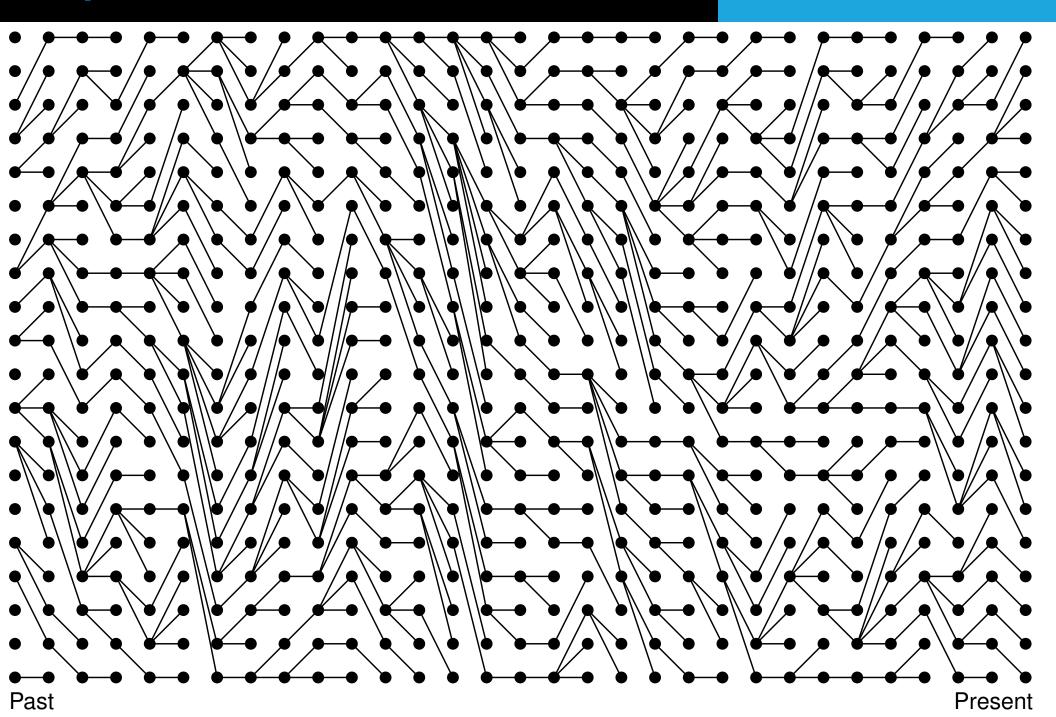
Past

- •
- •
- •
- •
- •
- •
- •
- •
- •
- •
- •
- •
- •
- •
- •
- •
- •
- •
- •

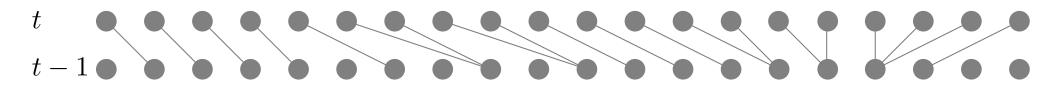
Past



Past

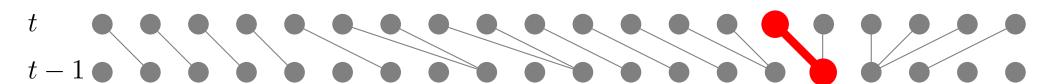


Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation t-1. If we assume that there are 2N chromosomes then the probability of sharing a common ancestor in the last generation is



Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation t-1. If we assume that there are 2N chromosomes then the probability of sharing a common ancestor in the last generation is

1.0



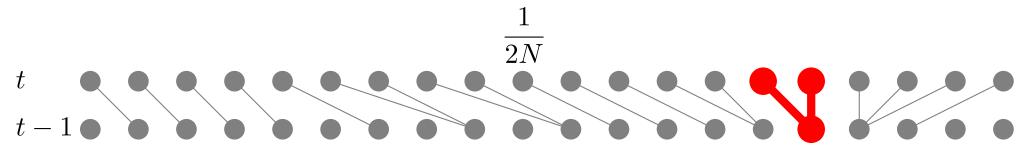
Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation t-1. If we assume that there are 2N chromosomes then the probability of sharing a common ancestor in the last generation is

$$t \\ t-1$$

Population model

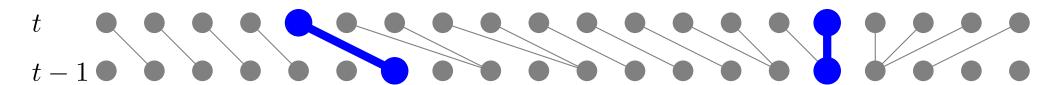
Wright

Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation t-1. If we assume that there are 2N chromosomes then the probability of sharing a common ancestor in last generation is



The probability that two randomly picked chromosome do not have a common ancestor is

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



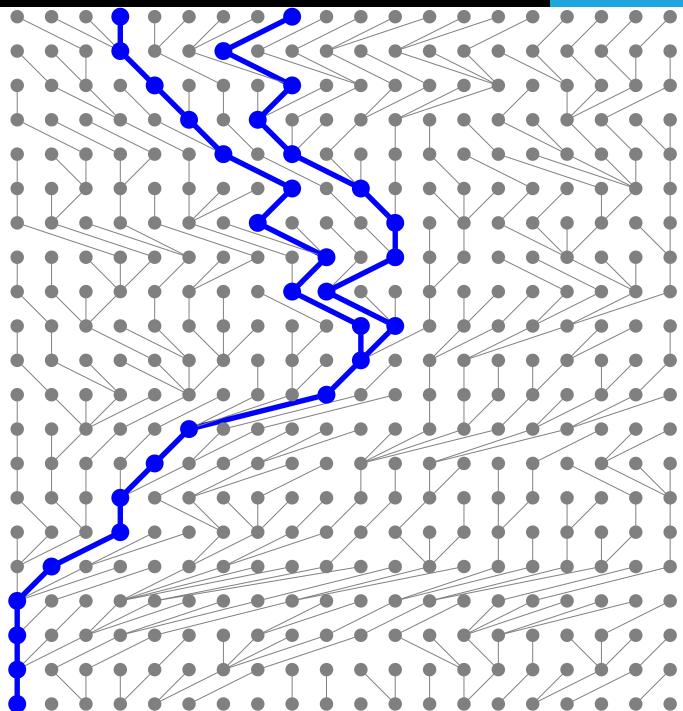
Population model

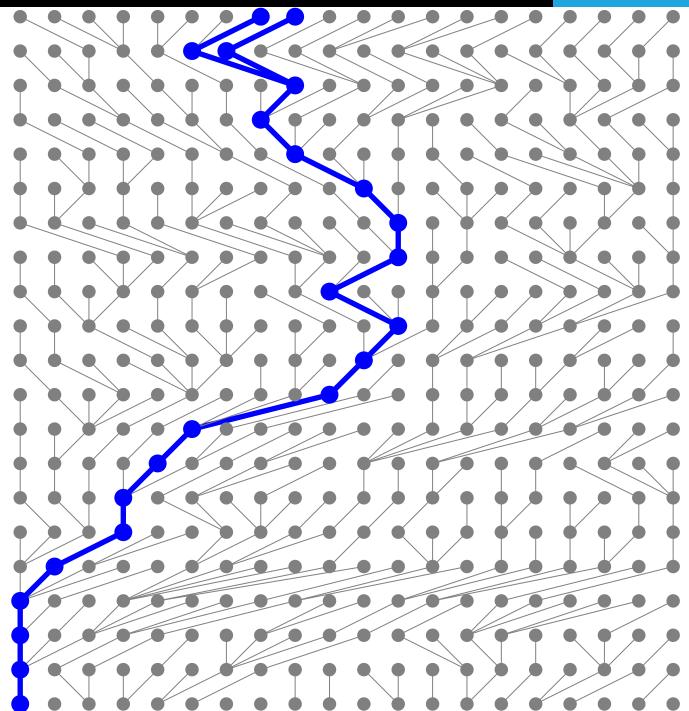
If we know the genealogy of the two individuals then we can calculate the probability as

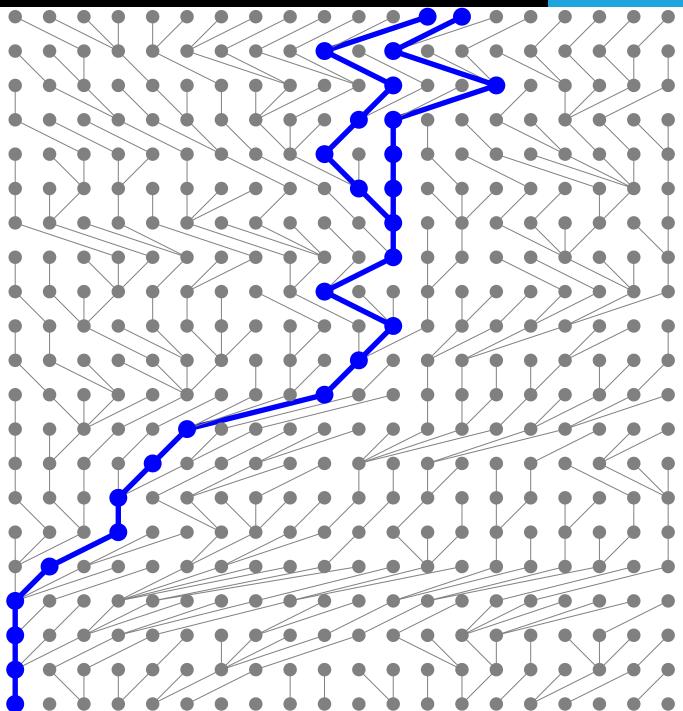
$$P(\tau|N) = \left(1 - \frac{1}{2N}\right)^{\tau} \left(\frac{1}{2N}\right)$$

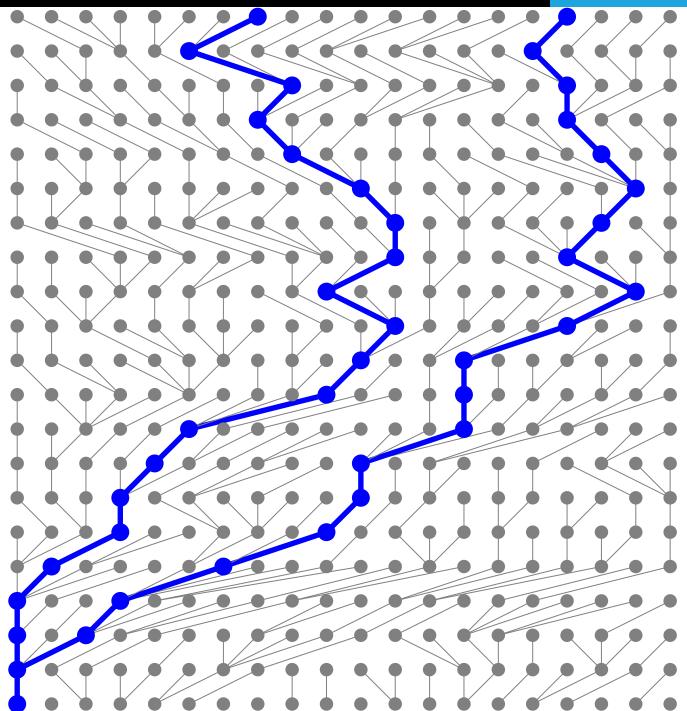
where τ is the number of generations with no coalescence. This formula is the Geometric Distribution and we can calculate the expectation of the waiting time until two random individuals coalesce:

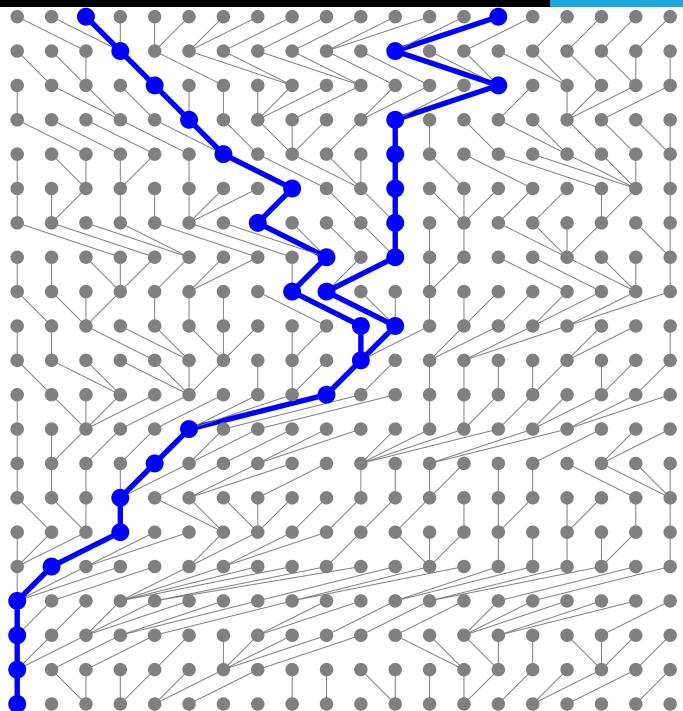
$$\mathbb{E}(\tau) = 2N$$

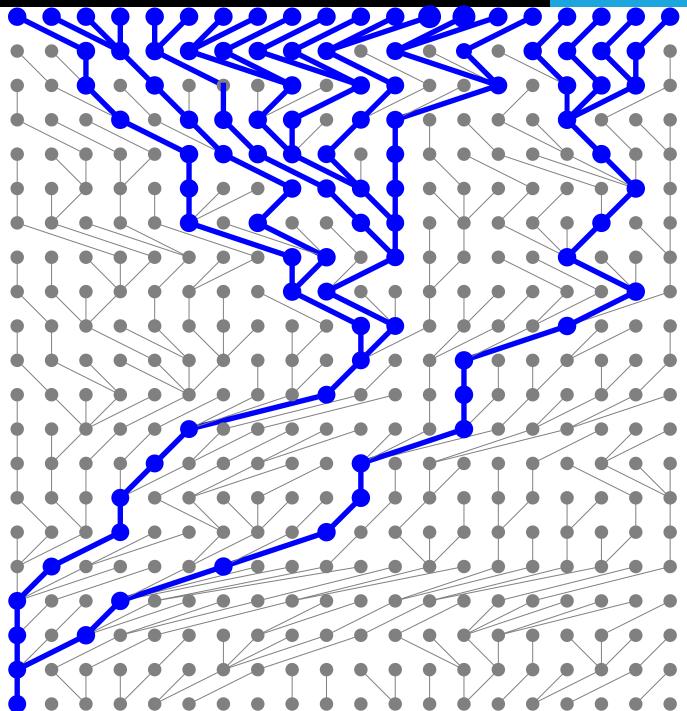


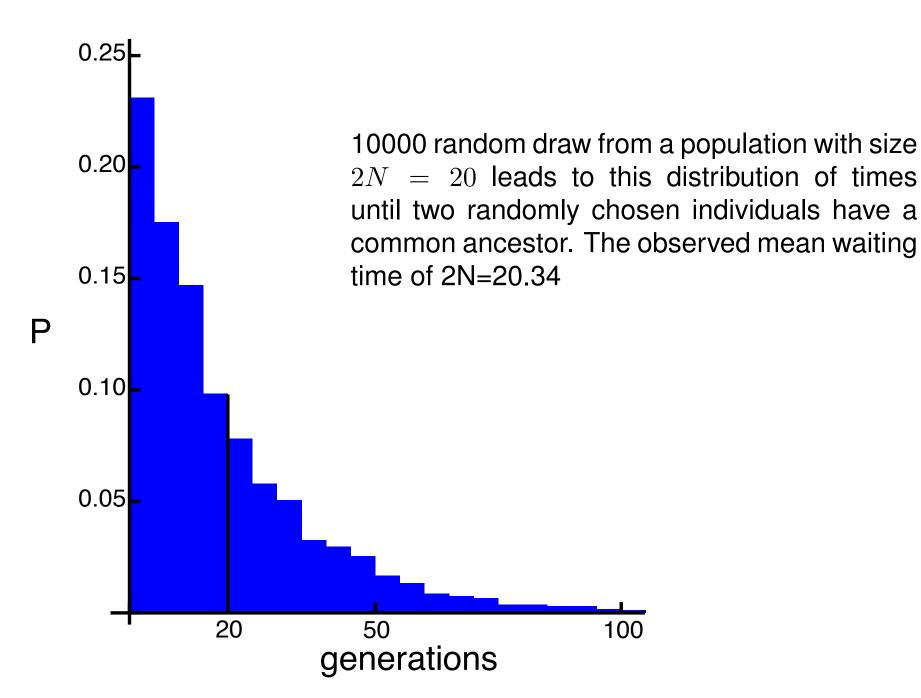








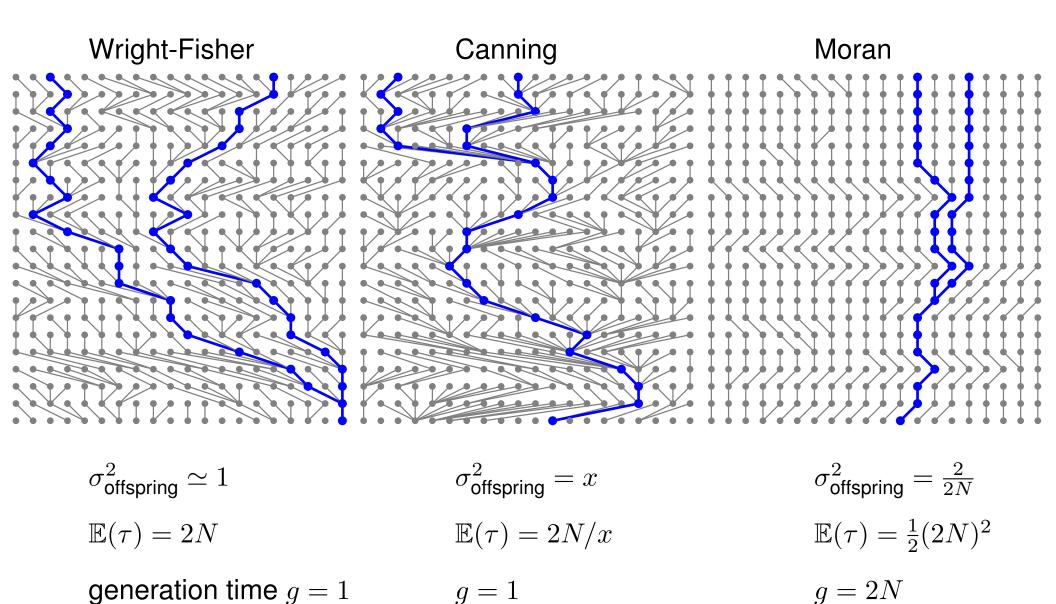




Observations

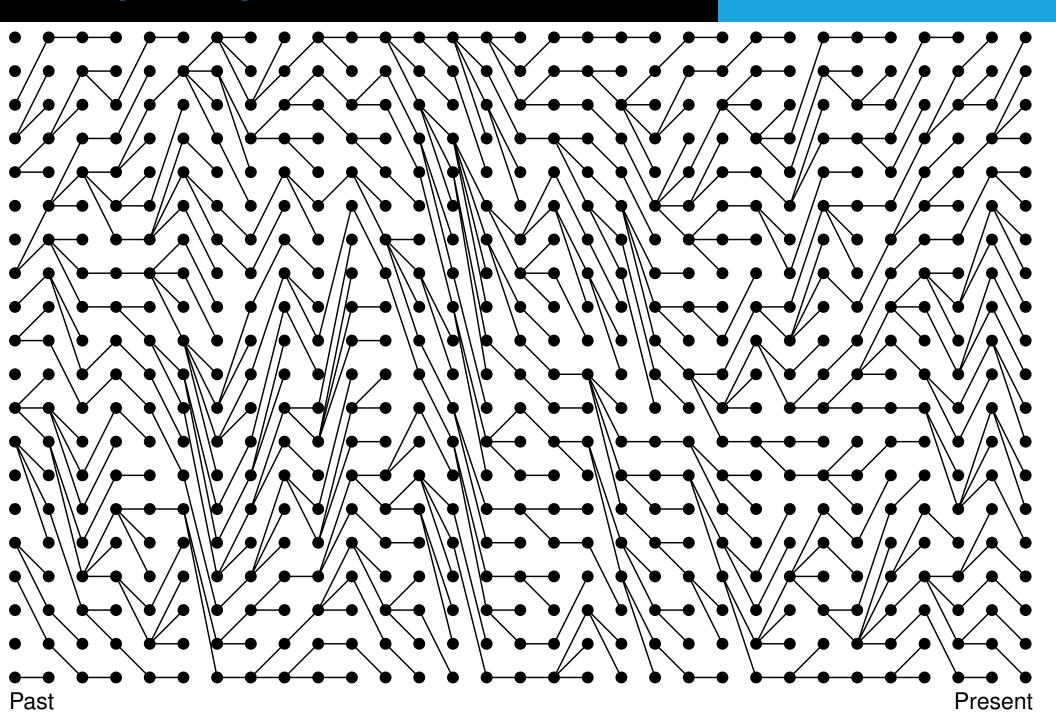
- For the time of coalescence in a sample of $\overline{\text{TWO}}$, we will wait on average 2N generations assuming it is a Wright-Fisher population
- The model assumes that the generations are discrete and non-overlapping
- Real populations do not necessarily behave like a Wright-Fisher (the 'ideal' population)
- We assume that calculation using Wright-Fisher populations can be extrapolated to real populations.

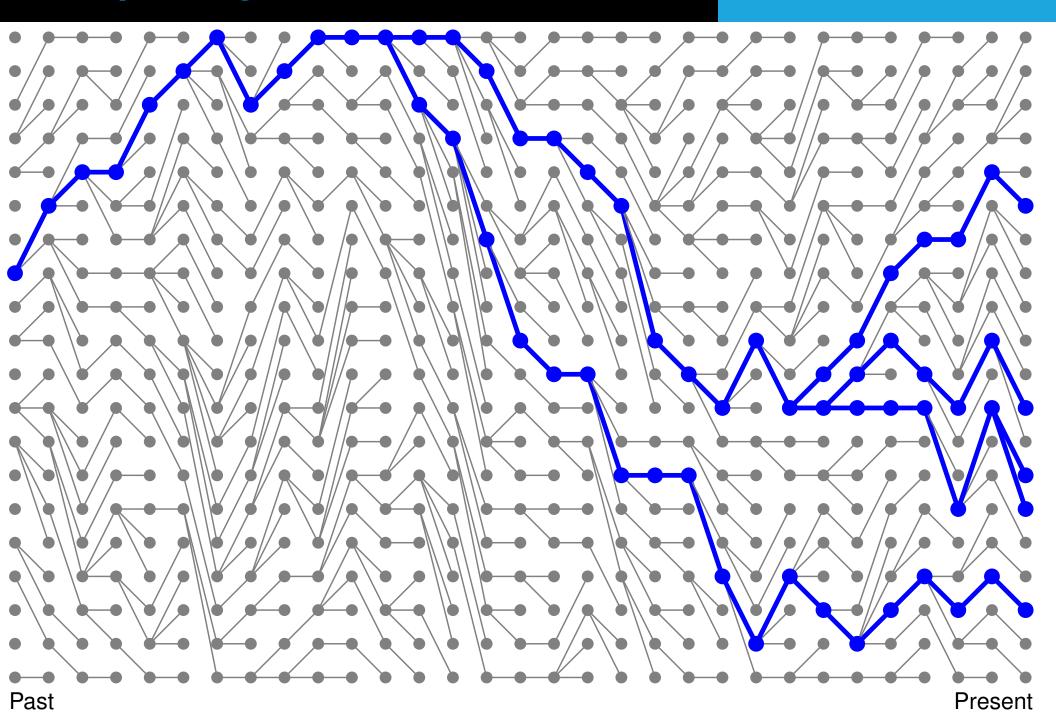
Other population models

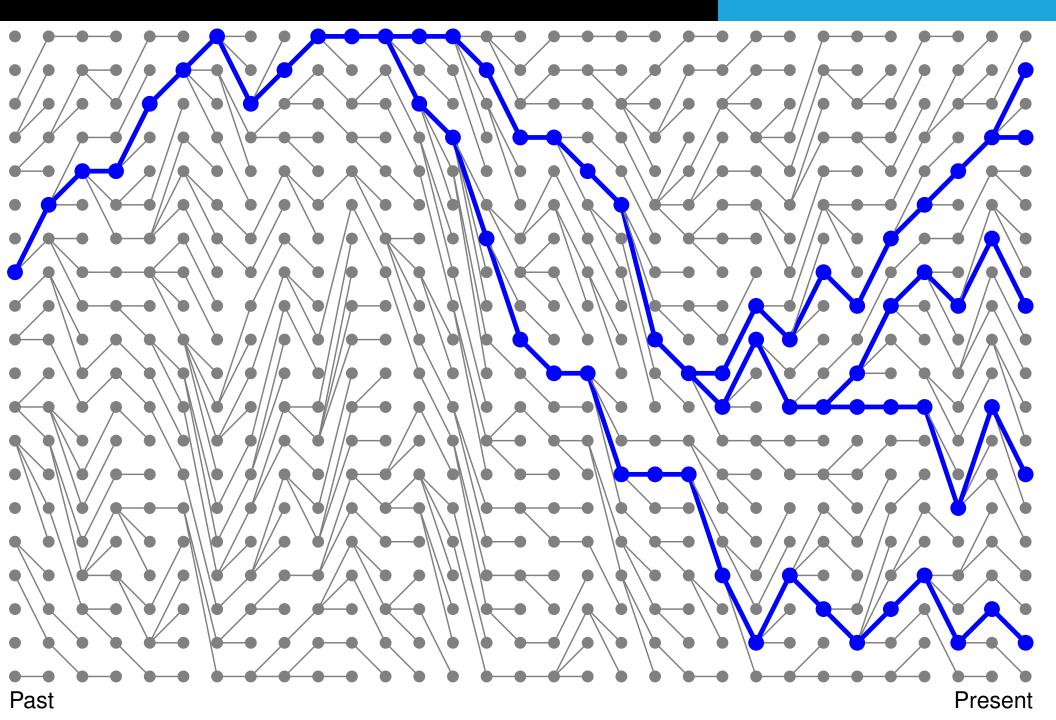


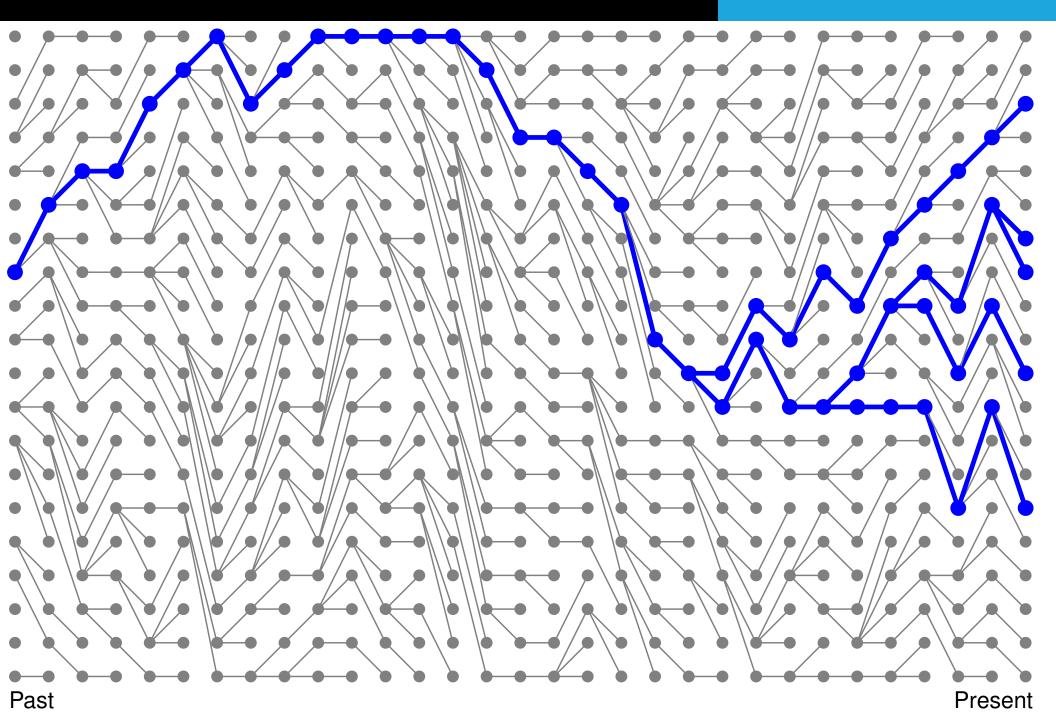
You can generate graphs like this using the python program *popsim* (check out my faculty page for the link)

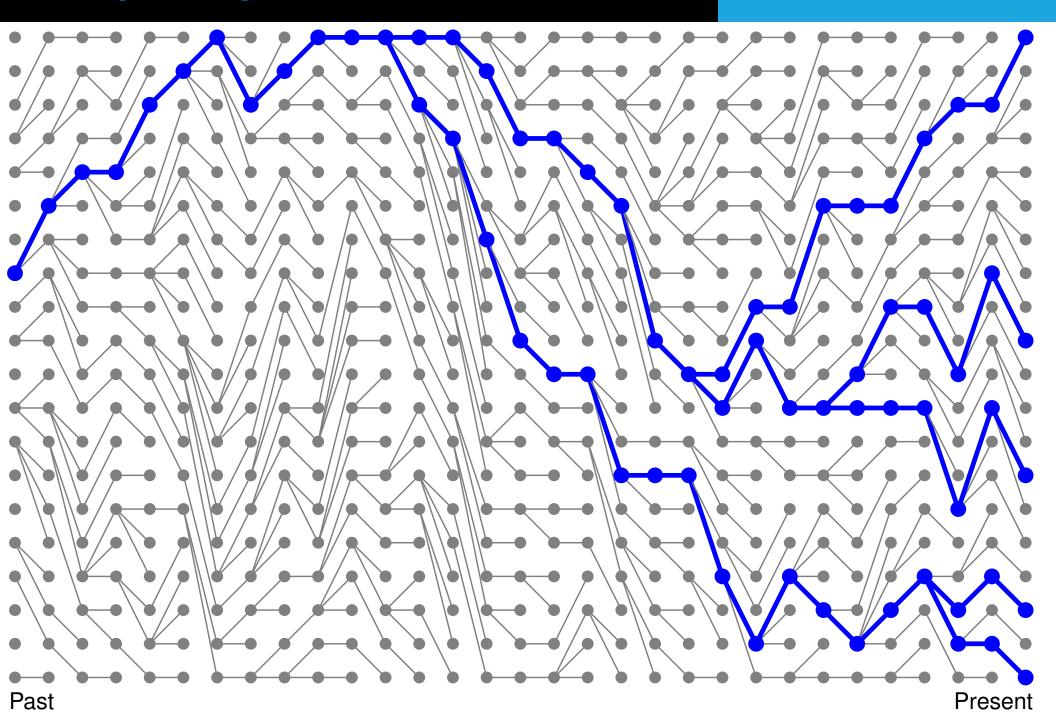
30 of 99 - ©2016 Peter Beerli











Samples larger than two

Sir J. F. C. Kingman described in 1982 the n-coalecent. He showed the behavior of a sample of size n, and its probability structure looking backwards in time.

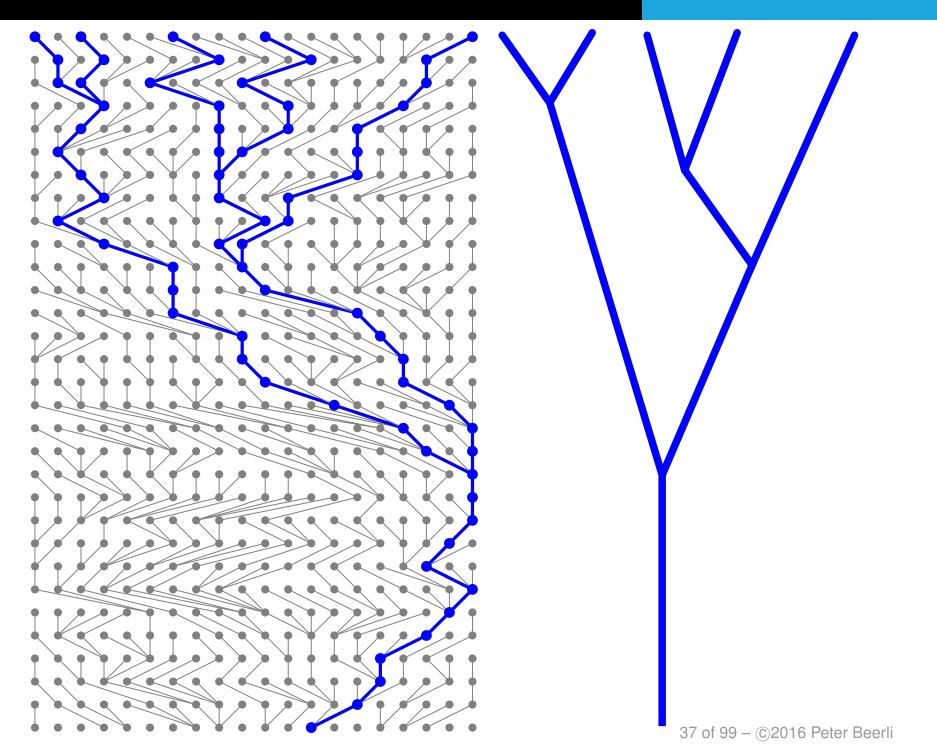
General findings:

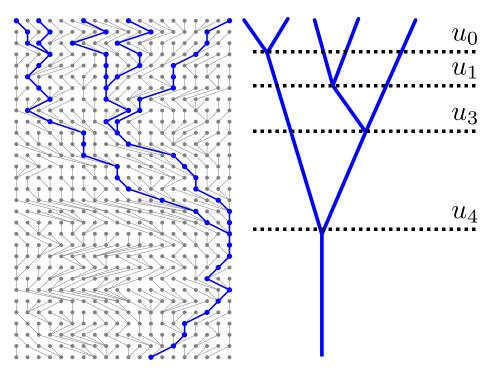
coalescence rate
$$= \binom{n}{2} = \frac{n(n-1)}{2}$$

Once a coalescence happened n is reduced to n-1 because two lineages merged into one. He then imposed a continuous approximation of the Canning's exchangeable model to get results.

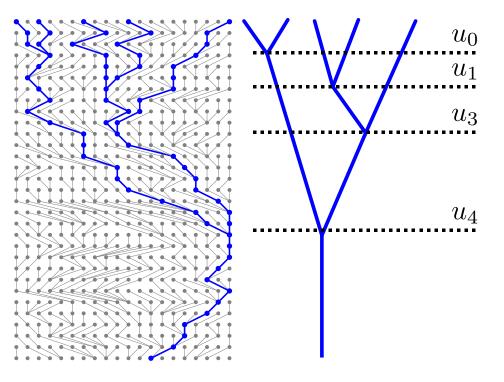






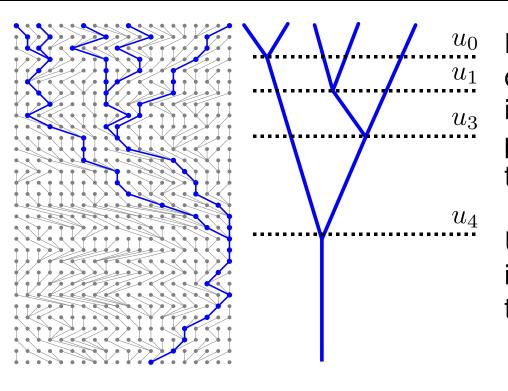


Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample n and the total population size N.



Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample n and the total population size N.

Using Kingman's coalescence rate and imposing a time scale we can approximate the process with a exponential distribution:



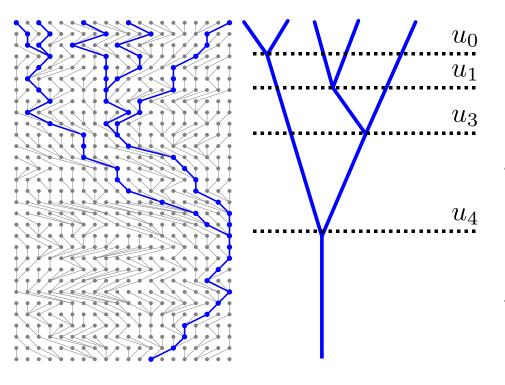
Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample n and the total population size N.

Using Kingman's coalescence rate and imposing a time scale we can approximate the process with an exponential distribution:

$$P(u_j|N) = e^{-u_j\lambda}\lambda$$

with the scaled coalescence rate

$$\lambda = \binom{k}{2} \frac{1}{2N} \times \text{Prob}(\text{others do not coalesce})$$



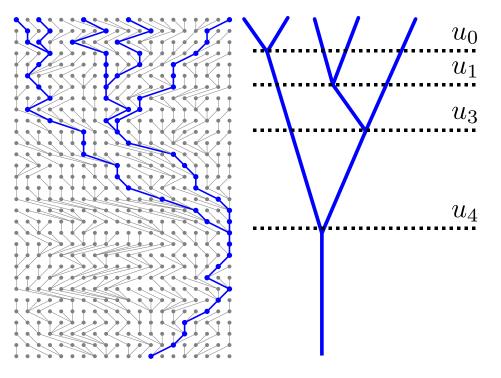
Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample n and the total population size N.

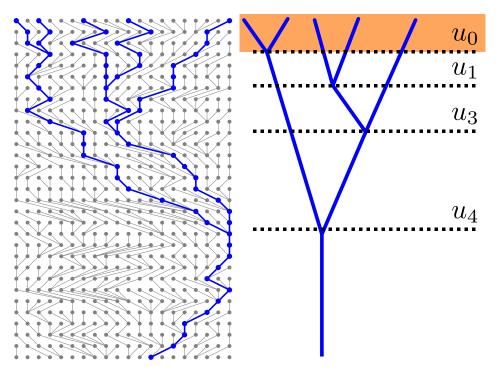
Using Kingman's coalescence rate and imposing a time scale we can approximate the process with a exponential distribution:

$$P(u_j|N) = e^{-u_j\lambda}\lambda$$

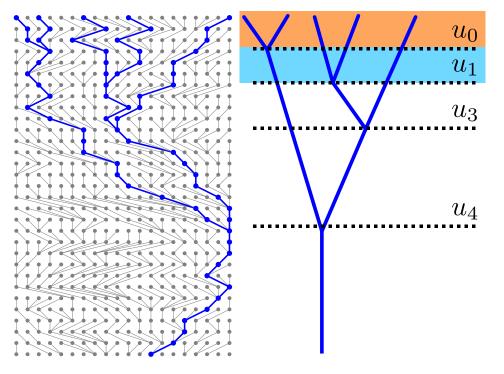
with the scaled coalescence rate

$$\lambda = \binom{k}{2} \frac{1}{2N} = \frac{k(k-1)}{2(2N)} = \frac{k(k-1)}{4N}$$



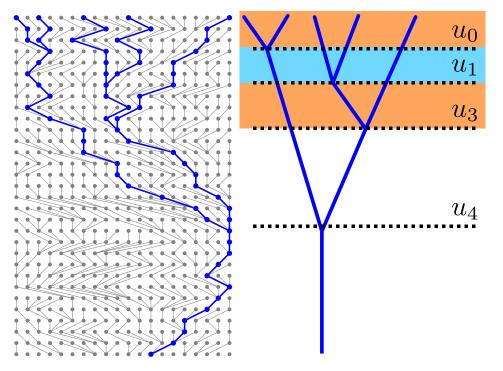


$$P(G|N) = P(u_0|N, i_1, i_2) \times$$

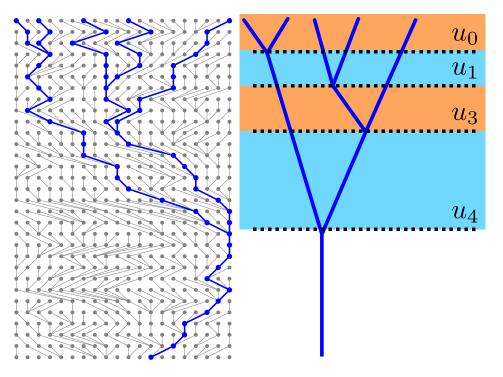


$$P(G|N) = P(u_0|N, i_1, i_2)$$

$$\times P(u_1|N, i_3, i_4)$$



$$P(G|N) = P(u_0|N, i_1, i_2)$$
 $\times P(u_1|N, i_3, i_4)$
 $\times P(u_3|N, i_{3,4}, i_5)$

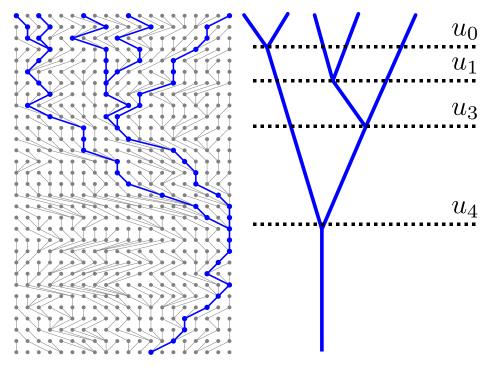


$$P(G|N) = P(u_0|N, i_1, i_2)$$

$$\times P(u_1|N, i_3, i_4)$$

$$\times P(u_3|N, i_{3,4}, i_5)$$

$$\times P(u_4|N, i_{1,2}, i_{3,4,5})$$



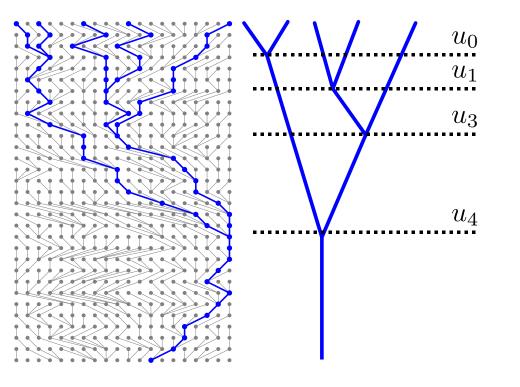
$$P(G|N) = P(u_0|N, i_1, i_2)$$

$$\times P(u_1|N, i_3, i_4)$$

$$\times P(u_3|N, i_{3,4}, i_5)$$

$$\times P(u_4|N, i_{1,2}, i_{3,4,5})$$

$$P(G|N) = \prod_{j=0}^{T} e^{-u_j \frac{k_j(k_j-1)}{4N}} \frac{2}{4N}$$



Each interval u_j is independent of the others, the expected length of the interval is the inverse of the coalescent rate. Thus we can sum these expectations to get to expectation of the depth of the genealogy.

$$\mathbb{E}(\tau_{\text{MRCA}}) = \text{Sum of the expectation of each time interval} = \sum_{j=0}^J \frac{4N}{k_j(k_j-1)}$$

$$\lim_{k\to\infty} \frac{\mathbb{E}(\tau_{\mathrm{MRCA}})}{\mathbb{E}(\tau_{\mathrm{MRCA}})} = 2N + \frac{2}{3}N + \frac{1}{3}N + \frac{1}{5}N + \frac{2}{15}N + \ldots = \frac{4N}{15} \qquad \lim_{k\to\infty} \sigma(\tau_{\mathrm{MRCA}}) = 4N$$

What is it good for?

If we know the genealogy G with certainty then we can calculate the population size N. Finding the maximum probability $\mathrm{P}(G|N,k)$ is simple, we evaluate all possible values for N and pick the value with the highest probability.

What is it good for?

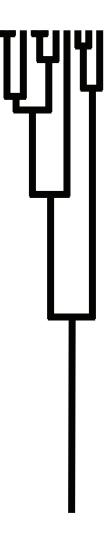
If we know the genealogy G with certainty then we can calculate the population size N. Finding the maximum probability P(G|N,k) is simple, we evaluate all possible values for N and pick the value with the highest probability.



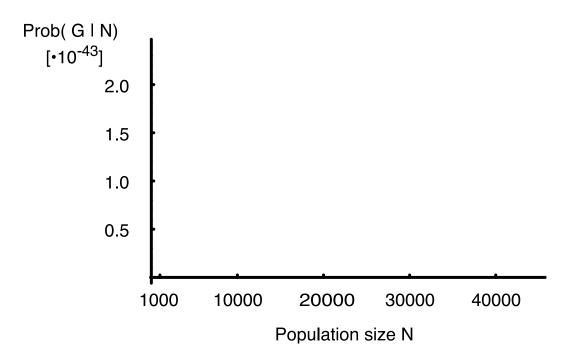
What is it good for?

If we know the genealogy G with certainty then we can calculate the population size N. Finding the maximum probability P(G|N,k) is simple, we evaluate all possible values for N and pick the value with the highest probability.



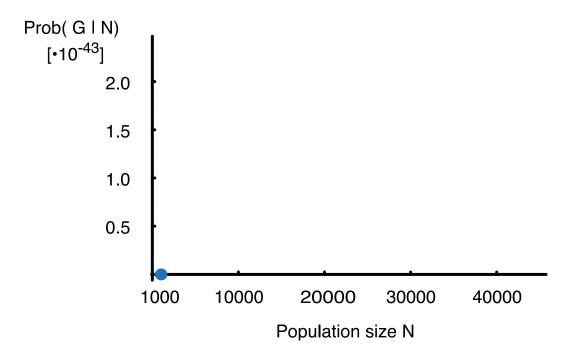






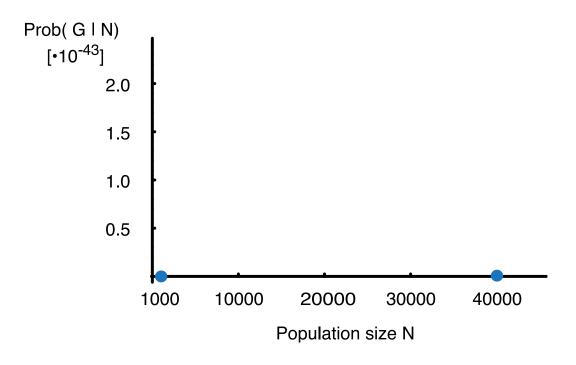
$$p(G|N,n) = \prod_{k=2}^{n} \exp\left(-u_k \frac{k(k-1)}{4N}\right) \frac{2}{4N}$$





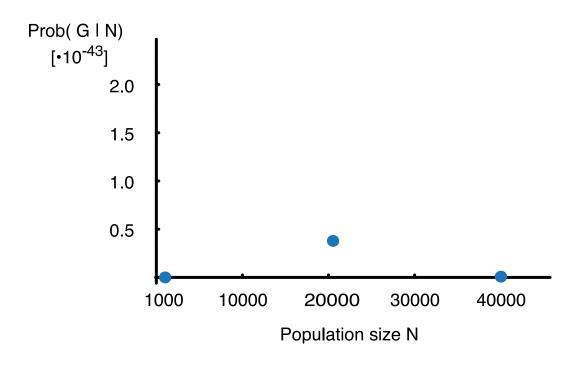
$$p(G|N,n) = \prod_{k=2}^{n} \exp\left(-u_k \frac{k(k-1)}{4N}\right) \frac{2}{4N}$$





$$p(G|N,n) = \prod_{k=2}^{n} \exp\left(-u_k \frac{k(k-1)}{4N}\right) \frac{2}{4N}$$



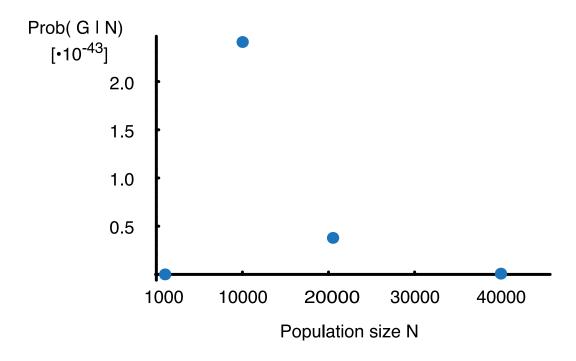


$$p(G|N,n) = \prod_{k=2}^{n} \exp\left(-u_k \frac{k(k-1)}{4N}\right) \frac{2}{4N}$$

using an oracle

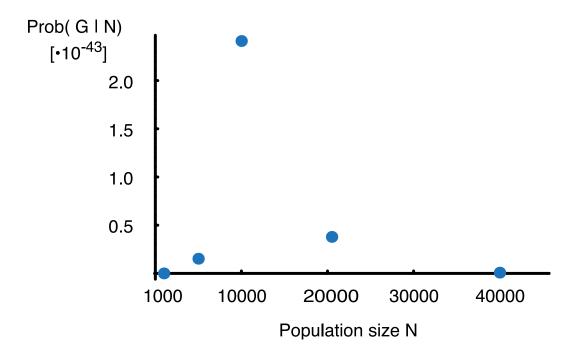
Population size estimation





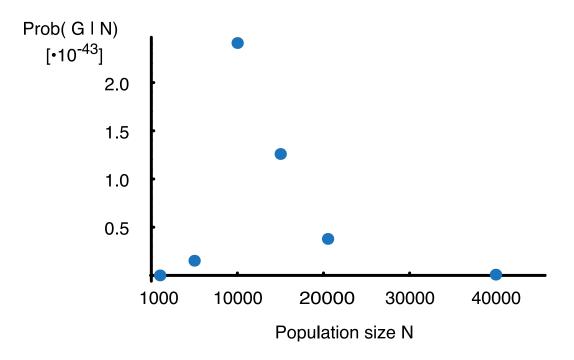
$$p(G|N,n) = \prod_{k=2}^{n} \exp\left(-u_k \frac{k(k-1)}{4N}\right) \frac{2}{4N}$$





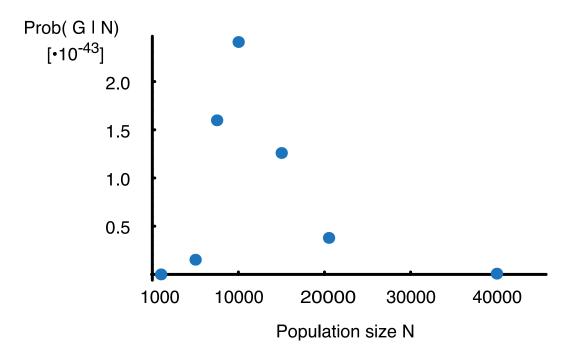
$$p(G|N,n) = \prod_{k=2}^{n} \exp\left(-u_k \frac{k(k-1)}{4N}\right) \frac{2}{4N}$$





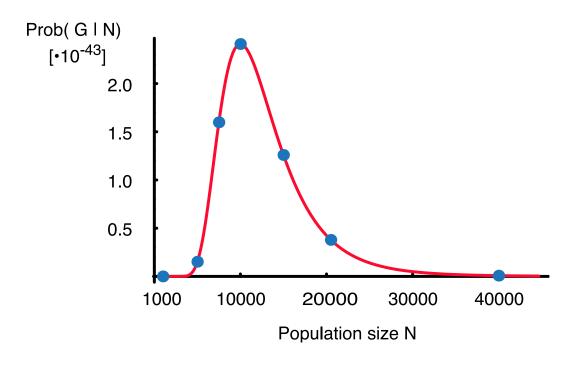
$$p(G|N,n) = \prod_{k=2}^{n} \exp\left(-u_k \frac{k(k-1)}{4N}\right) \frac{2}{4N}$$





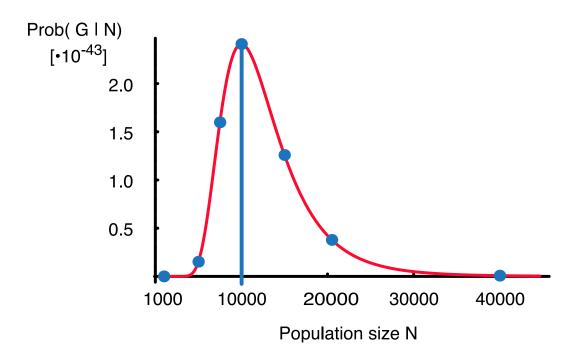
$$p(G|N,n) = \prod_{k=2}^{n} \exp\left(-u_k \frac{k(k-1)}{4N}\right) \frac{2}{4N}$$





$$p(G|N,n) = \prod_{k=2}^{n} \exp\left(-u_k \frac{k(k-1)}{4N}\right) \frac{2}{4N}$$



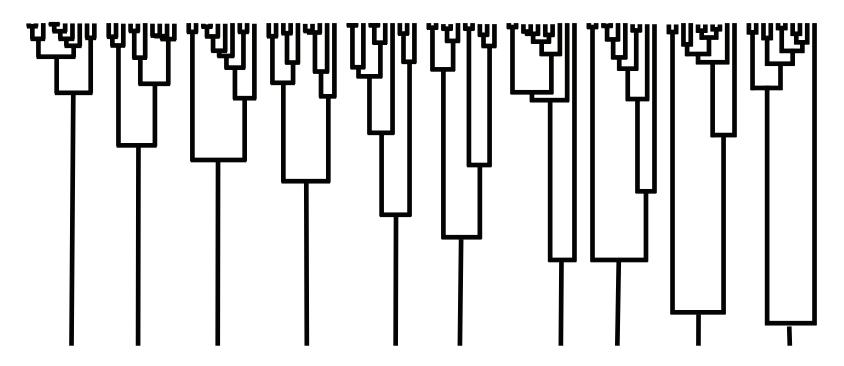


$$p(G|N,n) = \prod_{k=2}^{n} \exp\left(-u_k \frac{k(k-1)}{4N}\right) \frac{2}{4N}$$

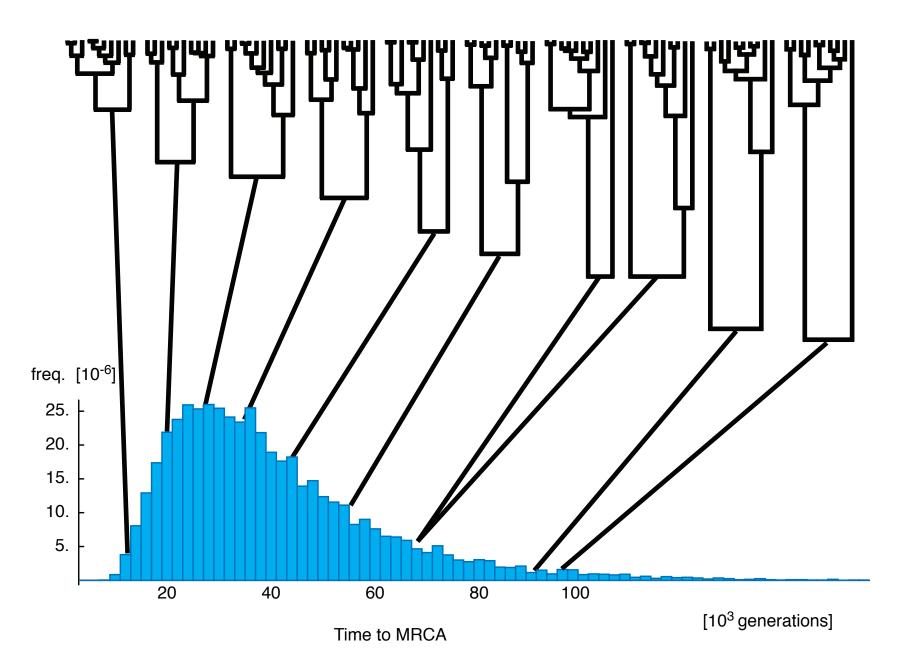
There are at least two problems with the oracle-approach:

- There is no oracle to gives us clear information!
- ◆ We do not record genealogies, our data are sequences, microsatellite loci!
- ♦ What about the variability of the coalescence process?





All genealogies were simulated with the same population size $N_e = 10,000$



MRCA = most recent common ancestor (last node in the genealogy)

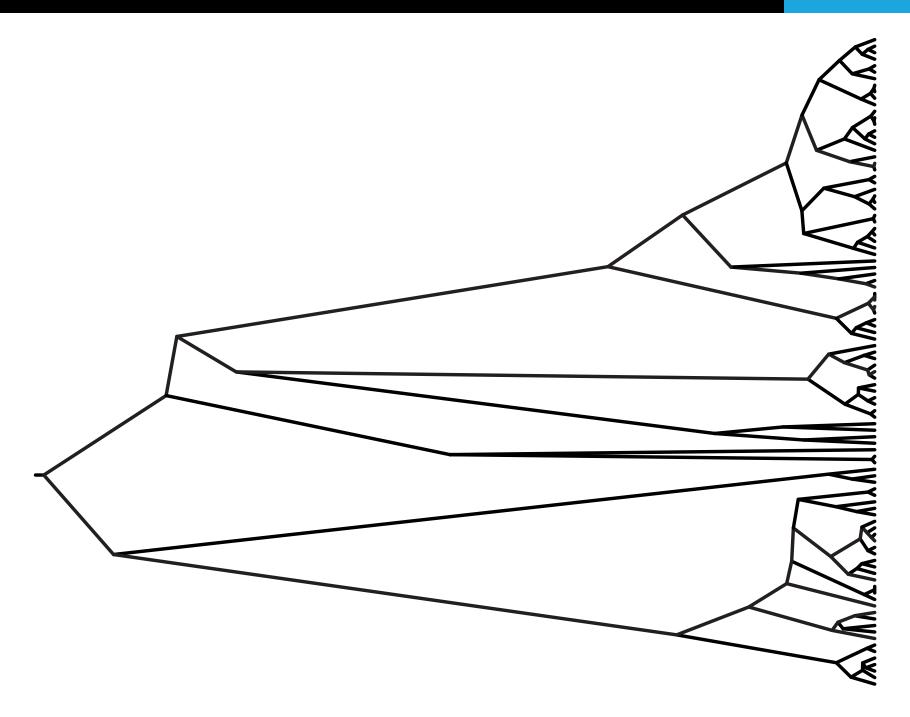
Kingman's n-coalescent is an approximation Sample Size

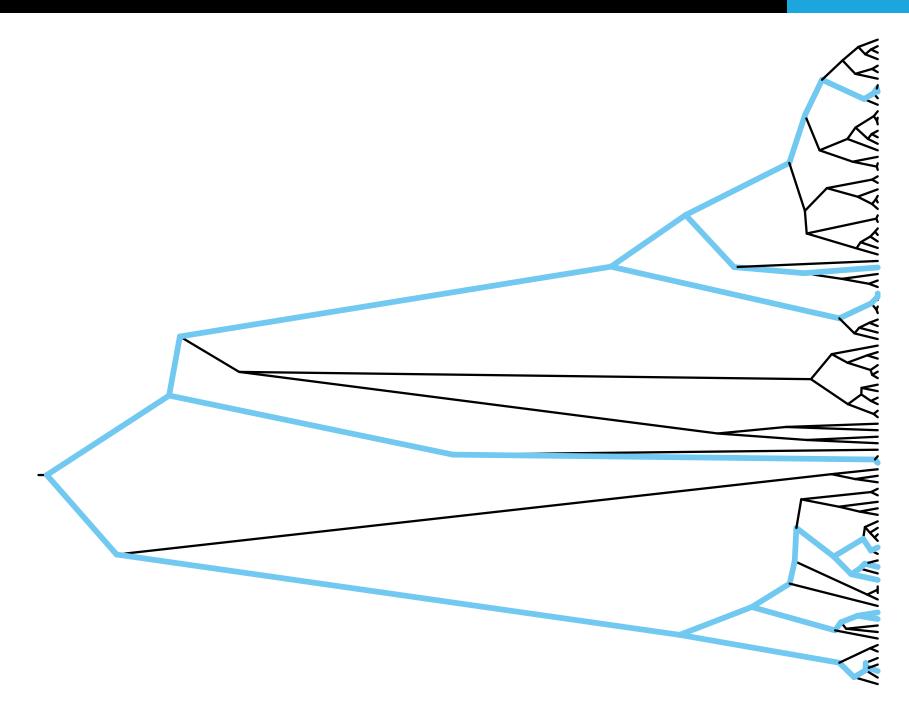
- All individuals have the same fitness (no selection).
- All individuals have the same chance to be in the sample (random sampling).
- The coalescent allows only merging two lineages per generation. restricts us to to have a much smaller sample size than the population size.

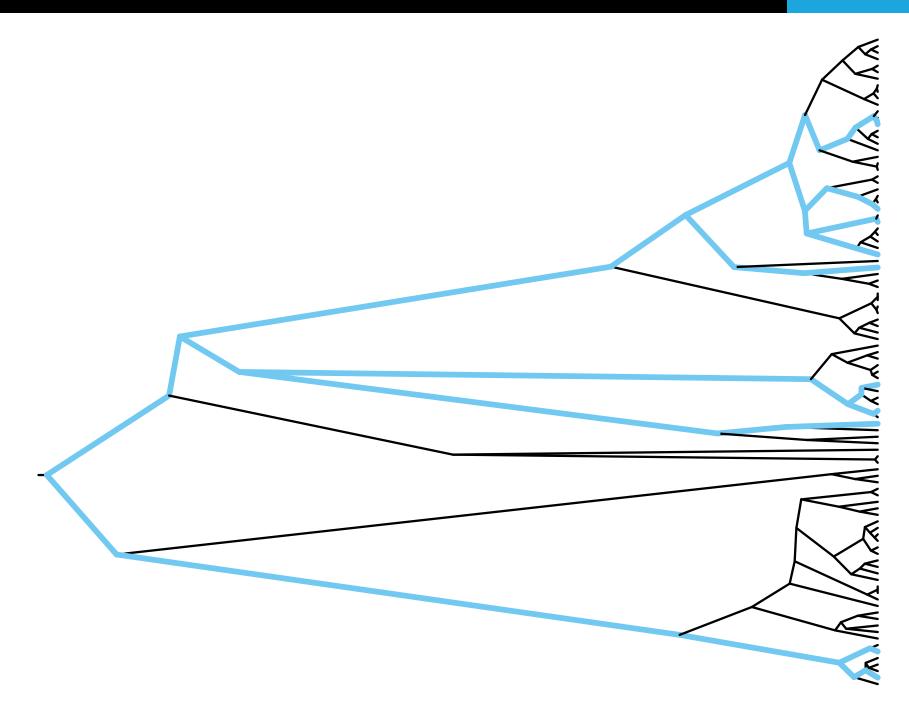
$$n \ll N$$

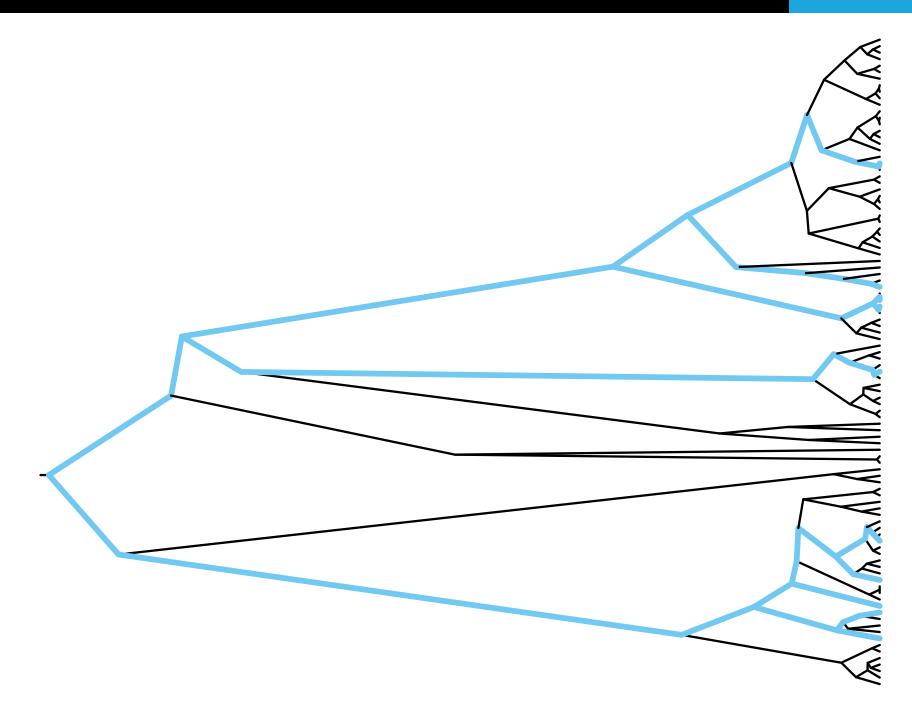
Yun-Xin Fu (2005) described the exact coalescent for the Wright-Fisher model and derived a maximal sample size $n < \sqrt{4N}$ for a diploid population. Although this may look like a severe restriction for the use of the coalescence in small populations, it turned out that the coalescence is rather robust and that even sample sizes close to the effective population size are not biasing immensely.

Kingman's n-coalescent is an approximation sample size









Observations

- lacktriangle Large samples coalesce on average in 4N generations.
- The time to the most recent common ancestor (TMRCA) has a large variance
- Even a sample with few individuals can most often recover the same TMRCA as a large sample.
- The sample size should be much smaller than the population size, although severe problems appear only with sample sizes of the same magnitude as the population size, or with non-random samples because Kingman's coalescence process assumes that maximally two sample lineages coalesce in any generation.
- With a known genealogy we can estimate the population size. Unfortunately, the true genealogy of a sample is rarely known.

Genealogy and data

our data looks like this:

rid2																				
rid3	ridl	GACTACE	AGCAC	GAAC	C	CECE	GGAG	AA <mark>G</mark> AG	ACGC	GA	G <mark>a</mark> GGGG		GALLAL	CAAD	G	AAA <mark>T</mark> AC	GEC	GGCG	CCC	AGCA
rid4		GACTAC	AGCAC	GAAC	I A I C	CECE	GGAG	aa <mark>g</mark> ag	ACGC	GA	G <mark>A</mark> GGGG	ATAT(GALA	CAA	GER C	AAATAC	GAC	GGCG	CCC	AGCA
rid5		GACTAC	AAGCAC	GAAC	A C	CTCT	GGAG	aa <mark>g</mark> ag	ACGC	GA	GAGGGG	ALAL (GALA	CAAL	GEAC	AAA <mark>T</mark> AC	GTAC	GGCG	CCC	AGCA
rid6		GACTAC	AGCAC	GAAC	C	CECE	GG AG	aa <mark>g</mark> ag	ACGC	AGA	GAGGGG	AL AL	GATA	CAAL	GEAC	AAA <mark>T</mark> AC	G	GGCG	CCC	AGCA
rid7	rid5	GACTAC	AGCAC	GAAC	E C	CEC	GGAG	aa <mark>g</mark> ag	ACGC	AGA	GAGGGG	ALL ALL	GAHAH	CAAL	GEAC	AAA <mark>T</mark> AC	GAC	GGCG	CCC	AGCA
rid8		GACTAC	AAGCAC	GAAC	C	CECE	GGAG	AAGAG	ACGC	AGA	GAGGGG	AHAH(GALLA	CAA	GEAC	AAA <mark>T</mark> AC	GEAC	GGCG	CCC	AGCA
ridloty2b Gac ac age can can call the color of can can call the ca	rid7	GACTAC	AAGCAC	GAAC	A C	CECE	GGAG	AAGAG	ACGC	AGA	GAGGGG	AHAH(GALLAL	CAAL	GEAC	AAAEAC	GTAC	GGCG	CCC	AGC AL
ridloty2b GACILCARGCAI GARC AT C.C. C. GGAGAGA GACCAT GAGAGACA GATA CATACAC AC CACAACAC GAGCACCCI GCA C.C. C. C. C. C. C. C. GGAGAGA GACCAT GAGGAGGAA GATA CATACAC CACAACAC GACCACCCI GCA C.C. C. C	rid8	GACTAC	AGCAC	GAAC	E C	CECE	GG AG	AAGAG	ACGC	AGA	GAGGGG	ALL ALL	GAHAL	CAAL	GEAC	AAA <mark>H</mark> AC	GLAC	GGCG	CCC	AGC A
ridloty2b GACILCARGCAI GARC AT C.C. C. GGAGAGA GACCAT GAGAGACA GATA CATACAC AC CACAACAC GAGCACCCI GCA C.C. C. C. C. C. C. C. GGAGAGA GACCAT GAGGAGGAA GATA CATACAC CACAACAC GACCACCCI GCA C.C. C. C	rid9ty1b	GACTAC	AGCAC	GAAC	E C	CEC	GGAG	AAGAG	ACGA	GA	GGGGGG	AHAH (GAHAL	CAAL	GEAC	AAA <mark>T</mark> AC	GAC	GGCG	CCC	AGCA
bed1 GACIRCA GARCAT GAAC ATTOCC C GAGAGA GACCAT CAGGGGGGAL AGAIT CAT CACAAIRC C AC GGGACCCIGCA CYP1 GACIRCA GCAR GARC GARC TATALOC C GAGAGA GACCAT CAGGGGGGAL GAIT CAT CACAAIRC C AC GGGACCCIGCA CYP2 GACIRCA GCAT GAAC TATALOC C GAGAGA GACCAT CAGGGGGGAL GAIT CAT CACAAIRC C AC GG GACCCIGCA CYP3 GACIRCA GCAT GAAC TATALOC C GAGAGA GACCAT CAGGGGGGAL GAIT CATALOCA CAAAIRC C AC GG GACCCIGCA CYP4 CACAACAA GCAT GAAC TATALOC C GAGAGA GACCAT CAGGGGGGAL GATALOCA CAAAIRC C AC GG GACCCIGCA CIWEST1 GACIRCA GCAC GAAC TATALOC C C GGAGAGA GACCAT CAGGGGGGAL GATALOCA CAAAIRC C AC GG GACCCIGCA CIWEST2 CACIRCA GCAC GAAC TATALOC C C GGAGAGA GACCCT GAAGGGGGAL GATALOCA CAAAIRC C AC GG GACCCIGCA CIWEST2 CACIRCA GCAC CAAGCAC GAAC TATALOC C C GGAGAGA GACCCT GAAGGGGGAL GATALOCA CAAAIRC C AC GGGGACCCIGCA CIWEST2 CACIRCA GACCAC GAAC TATALOC C C GGAGAAG GACCCT GAAGGGGGAL GATALOCA CAAAIRC C C AC GGGGACCCIGCA CIWEST2 CILEAST2 CACIRCA GCAC GAAC TATALOC C C GGAGAAG GACCCT GAAGGGGGAAL GATALOCA CAAAIRC C C AC GGGGACCCIGCA CICACAT CACAAIRC C CACAAAIRC C C AC GGGGACCCIGCA CAAIRC C C AC GACCACAT CACAAIRC C C AC GACCACCT GCA CILEAST2 CACIRCA GCACCAG CAAC TATALOC C GGAGAAG GACCCT GACAGGGGGAAL GATALOCA CAAAIRC C C AC GGGGACCCCIGCA CACAAIRC C CACAAIRC C C AC GACCACCT GCA CACAAIRC C C AC GACCACCT GCACCT GCACCACCT GCACCA	rid10ty2b	GACTAC	AAGCAC	GAAC	C	CECE	GGAG	AAGAG	A <mark>C</mark> GA	GA	GGGGGG	AHAH(GALLA	CAA	GEAC	AAA AC	GEAC	GGCG	CCC	AGCA:
Cyp2 Cyp3 CAC ACCARGAT GAAC TATTALC C GGAACA GACCAT CAGGGGGGATA GATA CAAC C ACAATAC G AC GGGAC CCCEAGCA Cyp4 GAC ACCAGGAT GAAC TATTALC C GGAGAGAGACCAT CAGGGGGGATA GATACAAC CACAATAC G AC GGGACCCCAGCA Cilwest1 GAC ACCAGGAT GAAC TATTALC C GGAGAGAGACCAT CAGGGGGATA GATACAAC CACAATAC G AC GGGACCCCAGCA Cilwest2 CACACAGCAT GAAC TATTACCC GGAGAGACACCCC CAGGGGGATA GATACAAC CACAATACC G AC GGGACCCCAGCA Cilwest2 CACACAGCAT GAAC TATTACCC GGAGAGAGACCCT CAGGGGGATA GATACAAC CACAATACCG AC GGGGACCCCAGCA Cilwest2 CACACAGCAT GAAC TATTACCC GGAGAGAGACCCT CAGGGGGATA GATACAAC CACAATACCG AC GGCGACCCCAGCA Cilwest3 CACACAGCAT GAAC TATTACCC GGGAAC GACCCT CAGGGGGATA GATACAAC CACAATACCG AC GGCGACCCCAGCA Cilwest4 CACACAGCAT GAAC TATTACCC GGGAAC GACCCT CAGGGGGATA GATACAAC CACAATACCG AC GGCGACCCCAGCA Cilwest5 CACACAGCAT GAAC TATTACCC GGGAAC GACCCT CAGGGGGATA GATACAAC CACAATACCG AC GGGGACCCCAGCAC Cilwest6 CACACAGCAT GAAC TATTACCC GGGAAC GACCAT CAGGGGGATA GATACAAC CACAATACCG AC GGGGGCACCCTAGCAC Cilwest7 CACACAGCAT GAAC TATTACCC GGGAACAGACCAT CAGGGGGATA GATACAAC CACAATACCG AC GGGGGCACCTAGCAC Cf.cara3 CACACAGCAT GAAC TATTACCC GGGAAGAGACAT CAGGGGGGATA GATACAAC CACAACACCCCTAGCACCACCACACACACACCACCACCACCACCACCACCAC	bed1	GACTAC	AAGC AZ	GAAC	A C	CECE	GGAG	AAGAG	A <mark>C</mark> GA	GA	GGGGGG	AHAH (GALLAL	CAAL	GEAC	AAAEAC	GTAC	GGCG	CCC	AGCAL
Cyp2 Cyp3 CAC ACCARGAT GAAC TATTALC C GGAACA GACCAT CAGGGGGGATA GATA CAAC C ACAATAC G AC GGGAC CCCEAGCA Cyp4 GAC ACCAGGAT GAAC TATTALC C GGAGAGAGACCAT CAGGGGGGATA GATACAAC CACAATAC G AC GGGACCCCAGCA Cilwest1 GAC ACCAGGAT GAAC TATTALC C GGAGAGAGACCAT CAGGGGGATA GATACAAC CACAATAC G AC GGGACCCCAGCA Cilwest2 CACACAGCAT GAAC TATTACCC GGAGAGACACCCC CAGGGGGATA GATACAAC CACAATACC G AC GGGACCCCAGCA Cilwest2 CACACAGCAT GAAC TATTACCC GGAGAGAGACCCT CAGGGGGATA GATACAAC CACAATACCG AC GGGGACCCCAGCA Cilwest2 CACACAGCAT GAAC TATTACCC GGAGAGAGACCCT CAGGGGGATA GATACAAC CACAATACCG AC GGCGACCCCAGCA Cilwest3 CACACAGCAT GAAC TATTACCC GGGAAC GACCCT CAGGGGGATA GATACAAC CACAATACCG AC GGCGACCCCAGCA Cilwest4 CACACAGCAT GAAC TATTACCC GGGAAC GACCCT CAGGGGGATA GATACAAC CACAATACCG AC GGCGACCCCAGCA Cilwest5 CACACAGCAT GAAC TATTACCC GGGAAC GACCCT CAGGGGGATA GATACAAC CACAATACCG AC GGGGACCCCAGCAC Cilwest6 CACACAGCAT GAAC TATTACCC GGGAAC GACCAT CAGGGGGATA GATACAAC CACAATACCG AC GGGGGCACCCTAGCAC Cilwest7 CACACAGCAT GAAC TATTACCC GGGAACAGACCAT CAGGGGGATA GATACAAC CACAATACCG AC GGGGGCACCTAGCAC Cf.cara3 CACACAGCAT GAAC TATTACCC GGGAAGAGACAT CAGGGGGGATA GATACAAC CACAACACCCCTAGCACCACCACACACACACCACCACCACCACCACCACCAC	cyp1	GACTAC	AGCA	GAAC		CECE	GG AG	AAGAG	A <mark>C</mark> GA	GA	GGGGGG	AHAH(GAHAL	CAAL	GEAC	AAA <mark>T</mark> AC	GEAC	GGTG	CCC	AGC A
cyp3 GACTACAR GARTAGATTATATTC GARGAGARA GARTAGATTATATC CARAGAGARA GARTAGARA GARTAGARTA		GACTACE	AGCA	GAAC	THE R	CECE	GGAG	aa <mark>g</mark> ag	A <mark>C</mark> GA	GA	GGGGGG	ALIAN (GATAT	CAAL	GEAC	AAATAC	GERC	GGGG	CCC	AGCA
cyp4 cilwest1 GACTACAGCAT GAACTTATACCC GGAGA GAGCGT GAGGGGATAT GATATCAAT GACAATAC GACTAC GG GACCCT AGCAT cilwest2 GACTACAGCAT GAACTTATCCC GGAGAAGAGGCT GAGGGGATAT GATATCAAT GACAAATAC GACTAC GGCGACCCT AGCAT cilwest2 GACTACAGCAT GAACTTATCCC GGAGAAGAGGCT GAGGGGGATATGATAT		GACTACA	AAGCA	GAAC	TATTA	CECE	GGAG	aa <mark>g</mark> ag	A <mark>C</mark> GA	GA.	GGGGGG	ATAT(GATA	CAA	GEL C	AAA <mark>T</mark> AC	GTAC	GGGG	CCC	AGCA
cilwest2 GACTACAGCAC GARCTTATEC CCC GGAGA GAGACCT GAGAGGGATAT GATATCAAT GACAATAC GACTAC GGCGACCT AGCACCIL GACTAC G		GACTAC	AAGCA	GAAC	TATTA	CTCT	GGAG	aa <mark>g</mark> ag	A <mark>C</mark> GA	GA.	GGGGGG	ALAL (GALA	CAAL	GLAC	AAA <mark>T</mark> AC	GTAC	GGGG	CCC	AGCAL
cileast1 GACTACAGCAC GARCTATECCC C GGAGARGAGGCT G GAGGGGATA GATACCAA GTACAATAC GACGAGCCCAGCACCCAGCACCCAGCACCCAGCACCCAGCACCCAGCACCCAGCACCCAGCACCCAGCACCCAGCACCCAGCACCCAGCACCCAGCACCCAGCACCCAGCACCCAGCACCCAGCACCCAGCACCCAGCACCAC	cilwestl	GACTAC	AGCAC	GAAC		CTCT	GG AG	aa <mark>g</mark> ag	ACGC	GA	GAGGGG	ALL ALL	GAHAL	CAAL	GEAC	AAA <mark>H</mark> AC	G	GGCG	CCC	AGC A
Cileast2 GACTACAGCATGAGCATGAGCATATACCCCGGAAAGAGAGAG	cilwest2	GACTAC	AGCAC	GAAC	E C	CEC	GGAG	aa <mark>g</mark> ag	ACGC	GA	GAGGGG	AL AL	GAHAH	CAAL	GEAC	AAA <mark>T</mark> AC	GEC	GGCG	CCC	AGCA
cf.cara2 cf.cara2 cf.cara3 cf.cara2 cf.cara3 cf.cara3 cf.cara4 cf.cara4 cf.cara4 cf.cara6 cf.cara5 cf.cara6 cf.cara6 cf.cara7 cf.cara7 cf.cara7 cf.cara7 cf.cara8 cf.cara7 cf.cara8 cf.cara8 cf.cara9 cf.car	cileast1	GACTAC	AAGCAC	GAAC	C	CECE	GGAG	AA <mark>G</mark> AG	ACGC	GA	GAGGGG	AHAH(GALA	CAA	GEAC	AAA <mark>T</mark> AC	GTAC	GGCG	CCC	AG <mark>C</mark> A
cf.cara2 cf.cara3 cf.car4 cf.car4 cf.car4 cf.car4 cf.car4 cf.car4 cf.car4 cf.car4 cf.car5 cf.car4 cf.car5 cf.car4 cf.car6 cf.car4 cf.car6 cf.car4 cf.car6 cf.car5 cf.car6 cf.car6 cf.car6 cf.car7 cf.car6 cf.car7 cf.car6 cf.car6 cf.car7 cf.car6 cf.car7 cf.car6 cf.car7 cf.car7 cf.car8 cr.car8 c	cileast2	GACTAC	AAGCAC	GAAC	A C	CECE	GG AG	aa <mark>g</mark> ag	ACGC	GA	GAGGGG	ALAL (GALA	CAAL	GEAC	AAAEAC	GTAC	GGCG	CCC	AGCA
cf.car4 cf.car5 cf.car4 cf.car5 cf.car6 cf.car6 cf.car6 cf.car7 cf.car7 cf.car6 cf.car7 cf.car7 cf.car8 cf.car8 cf.car8 cf.car8 cf.car8 cf.car8 cf.car9 cf.car8 cf.	cf.caral	GACTAC	AGCA	GAAC		CECE	GG AG	AAGAG	ACGA	GA	GGGGGG	ALA AL	GAHAL	CAAL	GEAR	AAA <mark>T</mark> AC	GGGC	GGGG	CC	AGMA
cf.cer1 GAC TACAAGCAT GAAC TATTAT C C GGAGAAGAGACGAT GAGGGGGGGATAT GATTAT CAAT G TATAAATAC G TACAAGCAT GACC TAGTAT CAAT G TATAAATAC G TATA	cf.cara2	GACTAC	AGC	GAAC	A A	CCC	GGAG	aa <mark>g</mark> ag	A <mark>CG</mark> A	GA	GGGGGG	ALL ALL	GAMA	CAA	GENERAL PROPERTY OF THE PROPER	AAA <mark>T</mark> AC	GGGC	GGGG	CC	AGTA
cf.cer1 GACTACAAGCAT GAACTATATCTC GGAGAAGAGAGGAT GAGGGGGGATATGATACATATACTA GTATAAATACT GTACTAGTACAAGCAT GAGGGGGGATATGATACTACTAGTACTAGTACTAGTACTAGTACTAGTACTAGTACTAGTACTAGTACTAGTACTAGTACTAGTACTACTAGTA	cf.cara3	GACTAC	AAGCA	GAAC	TATTA	CECE	GGAG	aa <mark>g</mark> ag	A <mark>C</mark> GA	GA	GGGGGG	AHAH(GATA	CAA	GHAHI	AAA <mark>T</mark> A <mark>C</mark>	GEC	GGGG	CC	AG <mark>T</mark> A
cf.cer2 GACTACAAGCAT GAACTATTATCCCGGAGAGAGAGACGAT GAGGGGGGATAT GATATCAAT GTATAAATACTGTACTG	cf.car4	GACTAC	AA <mark>GC</mark> A	GAAC	ALLA	CTCT	GGAG	aa <mark>g</mark> ag	A <mark>CG</mark> A	GA	GGGGGG	AEAE	GALA	CAAL	GEALL	AAAEAC	GLAC	GGGG	CC	AG A
cf.cer3 cf.cer4 cf.cer4 cf.cer4 cf.cer4 cf.bed1 cf.bed2 cf.bed2 cf.bed3 cf.cer4 cf.bed4 cf.bed4 cf.bed4 cf.bed5 cf.bed5 cf.bed5 cf.bed6 cf.bed6 cf.bed7 cf.bed7 cf.bed8 cf.bed8 cf.bed8 cf.bed8 cf.bed8 cf.bed9 cf.bed	cf.cerl	GACTACE	AGCA	GAAC	- A - A	CCC	GG AG	aa <mark>g</mark> ag	A <mark>C</mark> GA	G.A	GGGGGG		GAHAL	CAA	GEALL	aaa <mark>t</mark> ac	G	GG G	CC	AG A
cf.cer4 cf.bed1 cf.bed2 cf.bed2 cf.bed3 cf.bed3 cf.bed4 cf.bed4 cf.bed5 cf.bed4 cf.bed5 cf.bed5 cf.bed6 cf.bed6 cf.bed7 cf.bed8 cf.bed	cf.cer2	GACTAC	AGCA	GAAC	ALC: N	CCC	GGAG	AAGAG	ACGA	GA	GGGGGG	ALIAN (GAMA	CAA		AAA AC	G	GGGG	CC	AGTA
cf.bed1 GACTACAAGCATGAACTTATATCTCTGGAGAAGAGACGATGAGGGGGGGATATGATACAATGCTATAAATACTGTACTGGACCTTAGTATCT. cf.bed2 GACTACAAGCATGAACTTATATCTCTGGAGAAGAGACGATGAGGGGGGATATGATACAATGCTATAAATACTG		GACTAC	AAGCA	GAAC	TATTA	CECE	GGAG	aa <mark>g</mark> ag	A <mark>C</mark> GA	GA	GGGGGG	AHAH(GATA	CAA	G AY		G	GGGG	CC	AG A
cf.bed2 cf.bed3 cf.bed3 cf.bed3 cf.bed4 cf.bed4 cf.bed5 cf.bed5 cf.bed5 cf.bed6 cf.bed6 cf.bed7 cf.bed8 cf.bed		GACTAC	AAGC AS	GAAC	ATTA	CTCT	GGAG	aa <mark>g</mark> ag	A <mark>CG</mark> A	GA	GGGGGG	ALAL (GALA	CAA	GEALE?	A.A.A.E.A.C	G	GGGG	CC	AG A
Cf.bed3 Cf.bed4 Cf.bed5 Cf.bed5 Cf.bed6 Cf.bed6 Cf.bed6 Cf.bed7 Cf.bed7 Cf.bed8 Cf.bed8 Cf.bed8 Cf.bed8 Cf.bed8 Cf.bed8 Cf.bed7 Cf.bed8 Cf.bed		GACTAC	AGCA	GAAC	A A	CECE	GGAG	AAGAG	A <mark>C</mark> GA	G.A	GGGGGG	ALL ALL	GATA	CAA	BASI	AAA <mark>T</mark> AC	GEAC	GG G	CC	AG A
cf.bed5 GAC ACAGCAC GAAC A A C C GGAGAAGAGAGAGA		GACTAC		GAAC	ATTA	CECE	GGAG	aa <mark>g</mark> ag	ACGA	GA	GGGGGG	ALA	GATA	CAA		AAA <mark>T</mark> AC	G	GG	CC	AGTA
cf.bed5 GACTACAAGCAT GAACTA TATC C GGAGAAGAGACGAT GAGGGGGGGATATGATA CAATGTATAAATACTGTACTG		GACTACA	AGCAY	GAAC	TATTA	CECE	GGAG	aa <mark>g</mark> ag	A <mark>C</mark> GA	GA	GGGGGG	ATAT(GALA	CAA	STATE	AAATAC	GAC	GG G	CC	AGTA
cf.bed6 cf.bed7 cf.bed8 cf.bed8 cf.bed8 cf.caagcat gaactat at Ccc ggagaagagacgat gagggggatat gatat caat gtataaatac gtataaaatac gtataaatac gtataaatac gtataaatac gtataaatac gtataaatac gtata		GACTAC	AGCAC	GAAC	TATTA	CECE	GGAG	AAGAG	A <mark>C</mark> GA	GA	GGGGGG	A A	GALA	CAA	GAL	AAA AC	GTAC	GG G	CC	AG A
cf.bed7 cf.bed8 cf.bed8 cpe6-GR cpe7-GR cgacaccccccccccccccccccccccccccccccccc		GACTAC	AGCA	GAAC	A A	CECE	GGAG	AAGAG	A <mark>C</mark> GA	G.A.	GGGGGG	AL AL	GATA	CAA	SEAS	AAA <mark>T</mark> AC	G	GG G	CC	AG A
cf.bed8 GACTACAAGCATTGAACTTATTATCCC GGAGAAGAGACGATTGAGGAGAGAGAGAGAGAGAGAGAGA		GACTACE	AGCAC	GAAC	A. A	CCC	GGAG	aa <mark>g</mark> ag	ACGA	GA	GGGGGG	A A	GATA	CAA	STATE	AAA AC	GAC	GG G	CC	AGTA
cf.bed8 GACTACAAGCAT GAACTATATCC GGAGAAGAGACGAT GAGGGGGGGATAT GATATCAATGTATAAATACTGTACTG		GACTAC	AGCA	GAAC	TATTA	CECE	GGAG	aa <mark>g</mark> ag	A <mark>C</mark> GA	GA.	GGGGGG	ATAT(GALA	CAA	GIAT	AAA <mark>T</mark> AC	G	GGGG	CC	AG A
epe6-GR GACTACAAGCACTGAACTTATCCCCTGGAGAAGAGAGAGGGGGATATGATATCAATGTACAAATACTGTACAGGGGGGGCCCTAGCATGAAGAGAGAAGAGAAGAGACTATCTGTACTACTGTACTACTGTACTACTGTACTGTACTGTACTACTGTACTGTACTACTGTACTACTGTACTGTACTACTGTACTACTACTGTA	cf.bed8	GACTAC	AAGCA	GAAC	TATTA	CCC	GGAG	aa <mark>g</mark> ag	A <mark>C</mark> GA	GA	GGGGGG	AEAE						GG G	CC	AGTA
cre04a-GR GAC ACAGCACC A CC C GGGGAGGGGGGGGGGGG		GACTACE	AGCAC	GAAC	E C	CECE	GG.AG						GAHAL					GGCG	CCC	AGCA
cre04a-GR GAC ACAGCACC A CC C GGGGAGGGGGGGGGGGG		GACTACE	AGCAC	GAAC		CTCT	GG <mark>A</mark> G						GATA	CAAL	GEAC	AAA <mark>T</mark> AC	GAC	GGCG	ccc	AGCA
cre5-CR CACACACCACCACCACCACCACCACCACCACCACCACCA		GACTACE	AAGCAC	GAACC	CIATIC	CECE	GGAG				GGGGGG	ATAT	GATA	CAA	GELAC:	AAA <mark>T</mark> AC	GERC			AG <mark>C</mark> A
	cre5-GR		AAGCAC		CATTC	CTCT							GALA	CAAL	GEAC	AAA <mark>T</mark> AC	GTAC	GG G	CC	AGC AC

Genealogy and data

our data looks like this:



Genetic data and the coalescent

- Finite populations loose alleles due to genetic drift
- lacktriangle Mutation introduces new alleles into a population at rate μ
- With 2N chromosomes we can expect to see every generation $2N\mu$ new mutations. The population size N is positively correlated with the mutation rate μ .
- With genetic data sampled from several individuals we can use the mutational variability to estimate the population size.

Population size

The observed genetic variability

$$\mathcal{S} = f(N, \mu, n).$$

Different N and appropriate μ can give the same number of mutations. For example, for 100 loci sampled from 20 individuals with 1000bp each, we get :

\overline{N}	μ	$4N\mu$	\hat{S}	σ_S^2
1250	10^{-5}	0.05	153.95	16.25
12500	10^{-6}	0.05	152.89	16.05

Using genetic variability alone therefore does not allow to disentangle N and μ .

With multiple dated samples and known generation time we can estimate N and μ independently.

Mutation-scaled population size

By convention we express most results as the compound $N\mu$ and an inheritance scalar x, for simplicity we call this the mutation-scaled population size $\Theta = xN\mu,$

where μ is the mutation rate per generation and per site. With a mutation rate per locus we use θ .

- igoplus for diploids: $\Theta = 4N\mu$.
- igoplus for haploids: $\Theta = 2N\mu$.
- For mtDNA in diploids with strictly maternal inheritance this leads to $\Theta=2N_f\mu$, and if the sex ratio is 1:1 then $\Theta=N\mu$

Most real populations do not behave exactly like Wright-Fisher populations, therefore we subscript N and call it the effective population size N_e , and consider Θ the mutation-scaled EFFECTIVE population size.

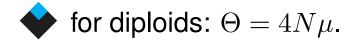
Mutation-scaled population size

By convention we express most results as the compound $N\mu$ and an inheritance scalar x, for simplicity we call this the mutation-scaled population size

 $\Theta = xN\mu$,

where μ is the mutation rate per generation and per site. With a mutation rate

per locus we use θ .







For mtDNA in diploids with strictly maternal inheritance this leads to $\Theta =$ $2N_f\mu$, and if the sex ratio is 1:1 then $\Theta=N\mu$

Most real populations do not behave exactly like Wright-Fisher populations, therefore we subscript N and call it the effective population size N_e , and consider ⊖ the mutation-scaled EFFECTIVE population size.

Historical humpback whale population size



Historical humpback whale population size

using the data by Joe Roman and Stephen R. Palumbi (Science 2003 301: 508-510)

$$\Theta = 2N_{
m Q}\mu$$

0.01529 Population size of the North Atlantic population, estimated using migrate

$$N_{
m Q}=rac{\Theta}{2\mu}$$

31,854 with $\mu = 2.0 \times 10^{-8} \mathrm{bp^{-1}year^{-1}}$ and a generation time of 12 years

$$N_e=N_{
m Q}+N_{
m C}$$

63,708 Sex ratio is 1:1

$$N_B = 2N_e$$

127,417 ratio N_B/N_e assumed, using other data

$$N_T = N_B rac{N_{
m juveniles} + N_{
m adults}}{N_{
m adults}}$$

203,867 from catch and survey data (used a ratio of 1.6)

More modern estimates for mtDNA: 150,000 [improved estimation of mutation rate]; for nucDNA: 112,000(45,000-235,000) [Conservation Genetics (2013) 14:103114]

Genetic data and the coalescent

Using the infinite sites model we use the number of variable sites S per locus to calculate the mutation-scaled population size:

$$\theta_W = \frac{S}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

from a sample of n individuals. For a single population the Watterson's estimator works marvelously well, but it is vulnerable to population structure.

Watterson's θ_W uses a mutation rate per locus! To compare with other work use mutation rate per site.

[in the second part of the coalescence talk I will discuss estimates based on site frequency spectra]

Construction of a versatile estimator

For Bayesian inference we want to calculate the probability of the model parameters given the data $p(\mathsf{model}|D)$.

Coalescent to describe the population genetic processes.

Mutation model to describe the change of genetic material over time.



Construction of a versatile estimator

We calculate the Posterior distribution $p(\Theta|D)$ using Bayes' rule

$$p(\Theta|D) = \frac{p(\Theta)p(D|\Theta)}{p(D)}$$

where $p(D|\Theta)$ is the likelihood of the parameters.



(almost) Felsenstein equation

$$p(D|\mathbf{\Theta}, G) = p(G|\mathbf{\Theta})p(D|G)$$

 $p(G|\mathbf{\Theta})$



The probability density of a genealogy given parameters.

p(D|G)



The probability density of the data for a given genealogy. Phylogeneticists know this as the tree-likelihood.

Felsenstein equation

$$p(D|\mathbf{\Theta}) = \int_G p(G|\mathbf{\Theta})p(D|G)dG$$

 $p(G|\mathbf{\Theta})$



The probability density of a genealogy given parameters.

p(D|G)



The probability density of the data for a given genealogy. Phylogeneticists know this as the tree-likelihood.

Felsenstein equation

$$p(D|\mathbf{\Theta}) = \sum_{G} p(G|\mathbf{\Theta})p(D|G)$$

 $p(G|\mathbf{\Theta})$



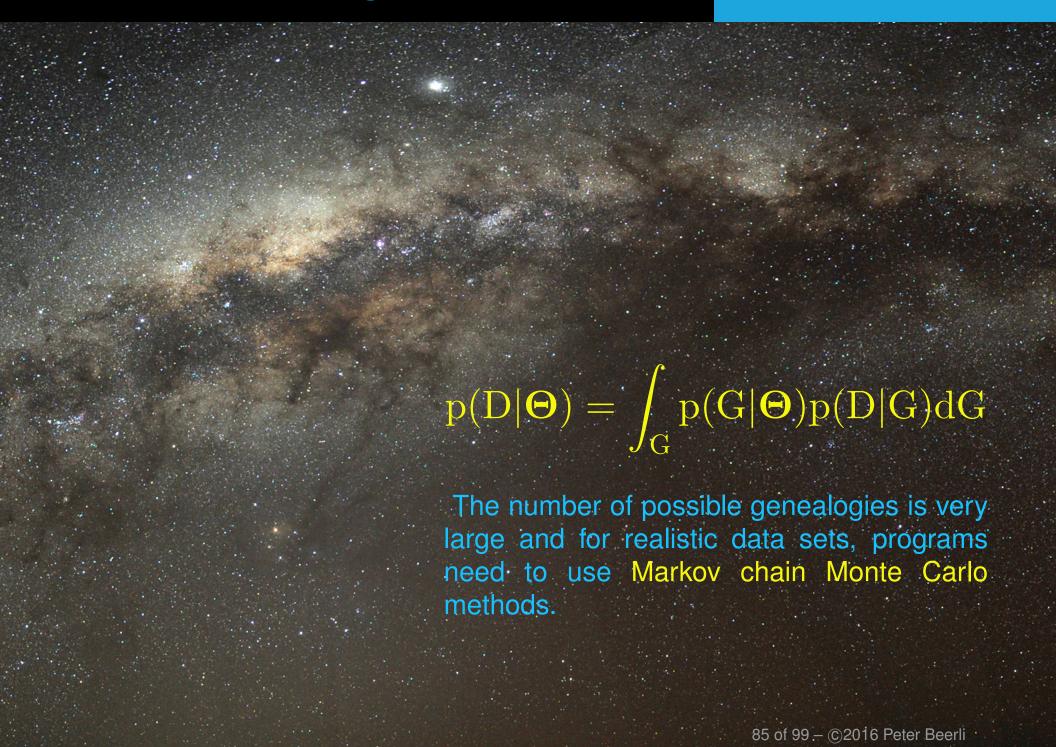
The probability of a genealogy given parameters.

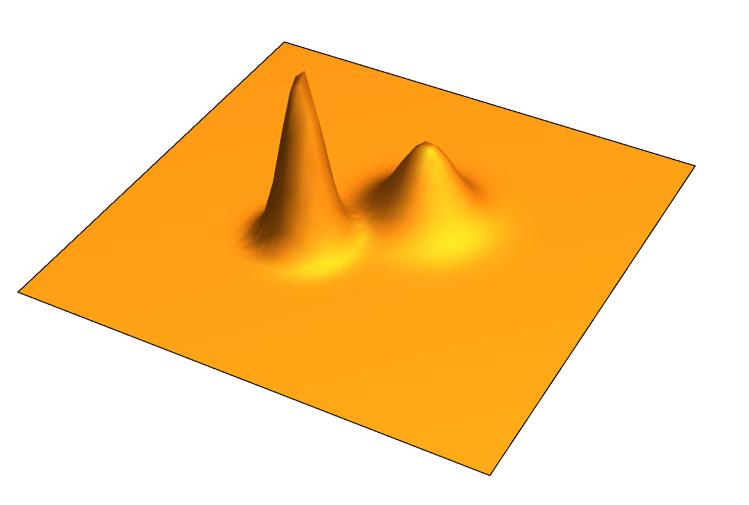
p(D|G)

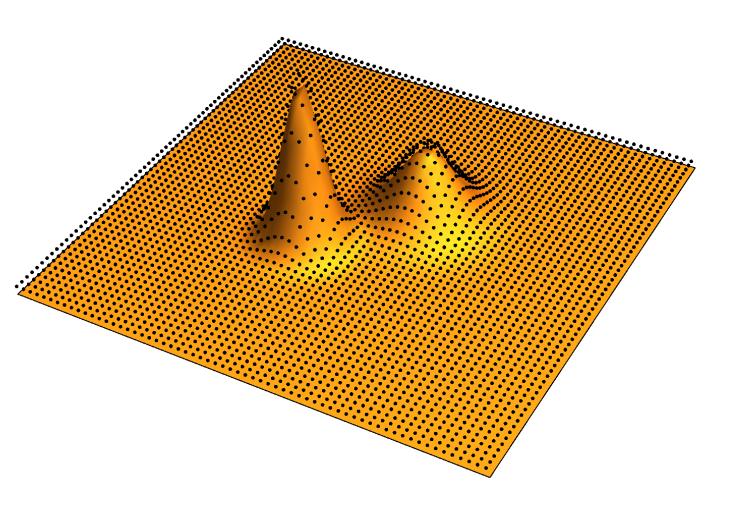


The probability of the data for a given genealogy. Phylogeneticists know this as the tree-likelihood.

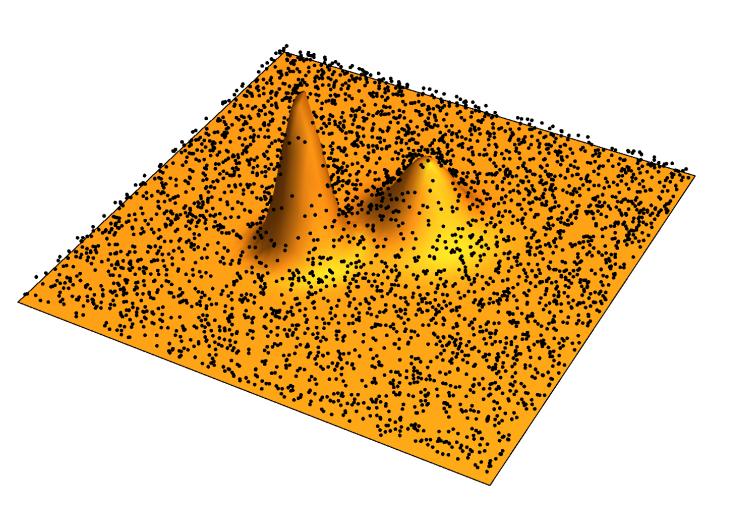
Problem with integration formula



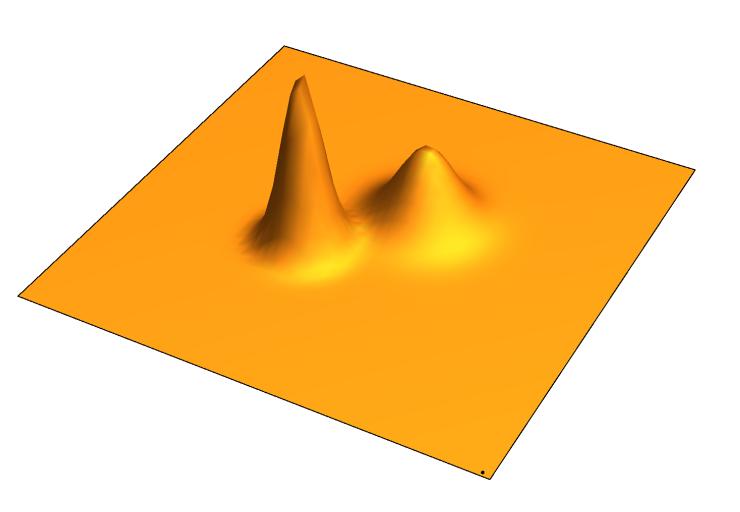




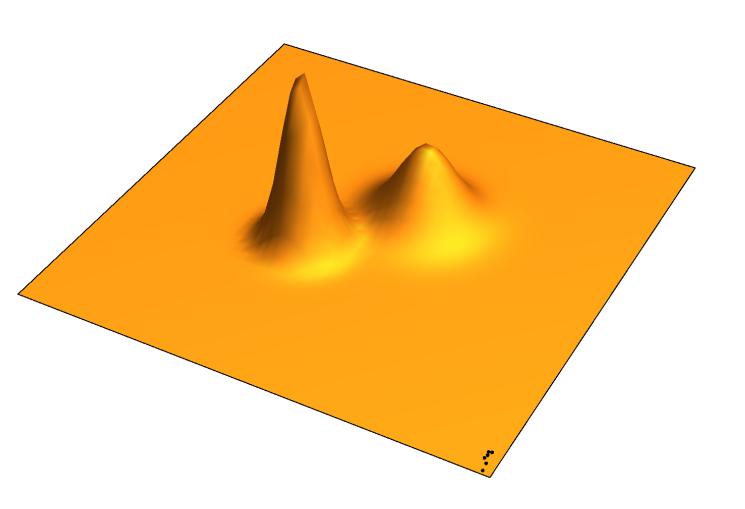




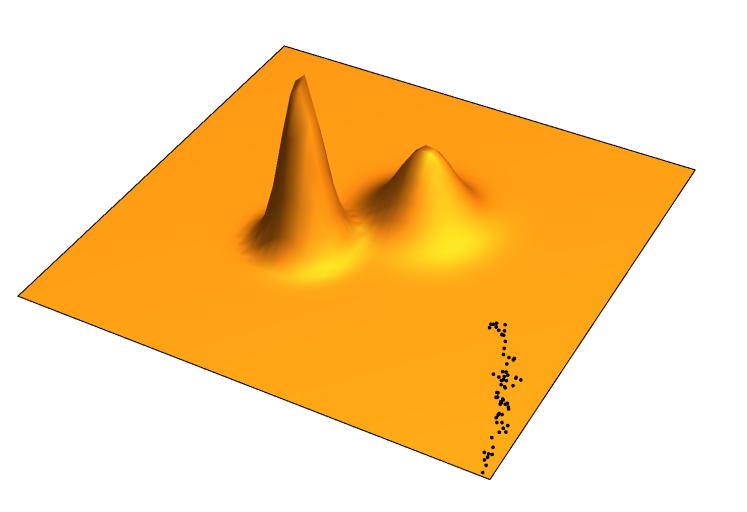




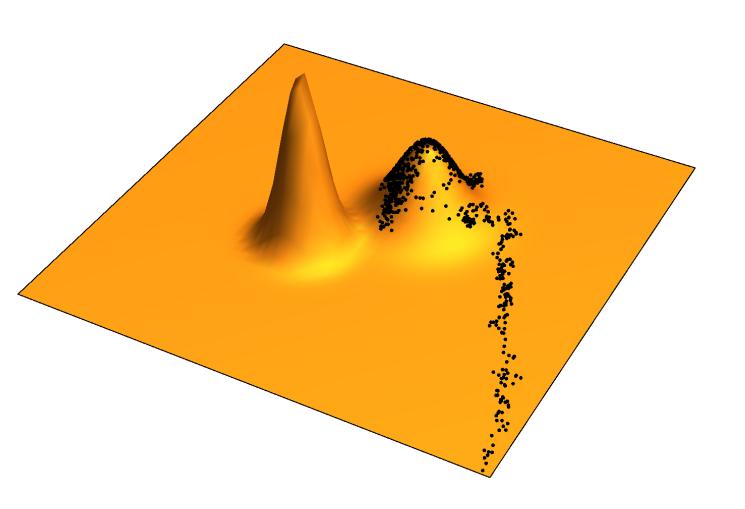




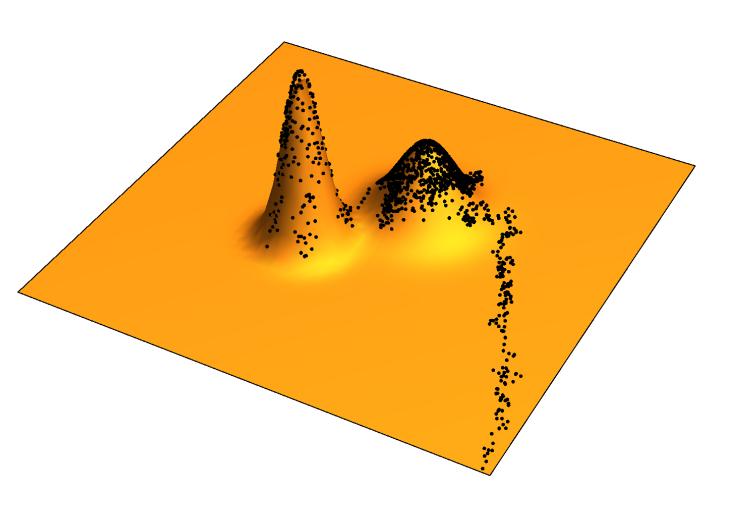




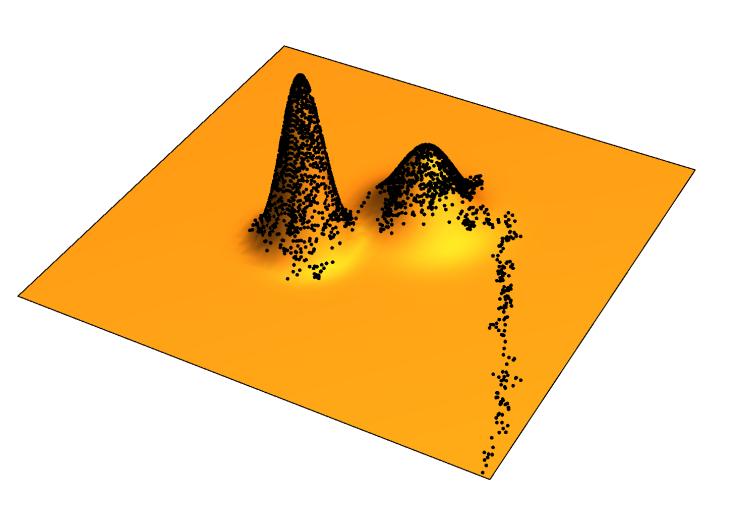


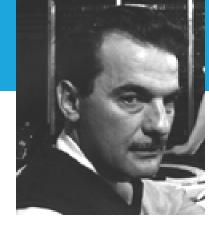


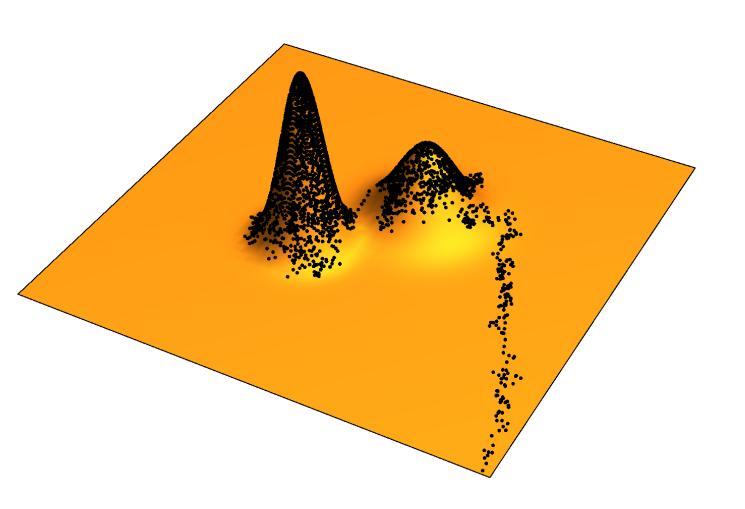














Inference of population size

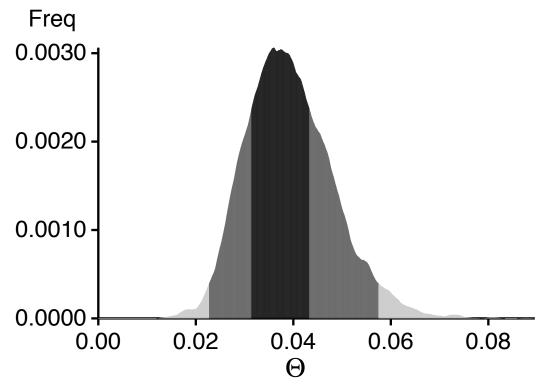


Proc. Natl. Acad. Sci. USA Vol. 88, pp. 8720-8724, October 1991 Evolution

Extensive mitochondrial diversity within a single Amerindian tribe (population genetics/molecular anthropology/Pacific Northwest/human evolution)

R. H. WARD*, BARBARA L. FRAZIER*, KERRY DEW-JAGER*, AND SVANTE PÄÄBO[†]
*Department of Human Genetics, School of Medicine, University of Utah, Salt Lake City, UT 84132; and [†]Department of Zoology, University of Munich, Luisenstrasse 14, D-8000 Munich 2, Federal Republic of Germany

[The Nuu-Cha-Nulth are organized in 14 nations totaling 8147 (Nuuchahnulth tribal council Indian registry from February 2006)]



Bayesian inference: $\Theta = 0.036$

Ward *et al* calculated $\Theta_{Ewens} = 0.043$

With a mutation rate of 0.32/site/million year and a generation time of 27 years we get $N_{\rm women}=2082$. Assuming same numbers of men and women and on average 2 children we get N=8328.

Inference of population size

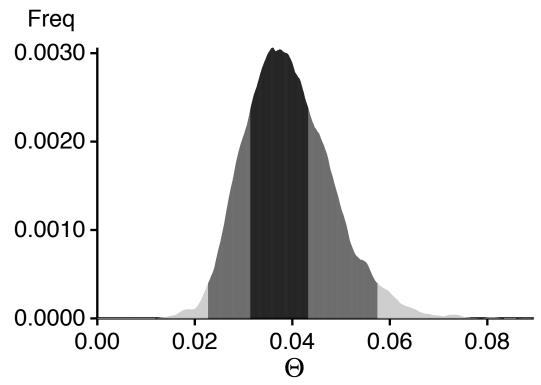


Proc. Natl. Acad. Sci. USA Vol. 88, pp. 8720-8724, October 1991 Evolution

Extensive mitochondrial diversity within a single Amerindian tribe (population genetics/molecular anthropology/Pacific Northwest/human evolution)

R. H. WARD*, BARBARA L. FRAZIER*, KERRY DEW-JAGER*, AND SVANTE PÄÄBO[†]
*Department of Human Genetics, School of Medicine, University of Utah, Salt Lake City, UT 84132; and [†]Department of Zoology, University of Munich, Luisenstrasse 14, D-8000 Munich 2, Federal Republic of Germany

[The Nuu-Cha-Nulth are organized in 14 nations totaling 8147 (Nuuchahnulth tribal council Indian registry from February 2006)]



Bayesian inference: $\Theta = 0.036$

Ward *et al* calculated $\Theta_{Ewens} = 0.043$

With a mutation rate of 0.32/site/million year and a generation time of 27 years we get $N_{\rm women}=2082$. Assuming same numbers of men and women and on average 2 children we get N=8328.

This sounds very good, but the 95%credibility interval is 5000 - 13500

References

Coalescent:

- Nuu-Cha-Nulth population size: J. Felsenstein. 1971. Inbreeding and variance effective numbers in populations with overlapping generations. Genetics 68:581-597;
- R. H. Ward, B. L. Frazier, Kerry Dew-Jager, and S. Pääbo. 1991. Extensive mitochondrial diversity within a single Amerindian tribe. PNAS 88:8780-8724;
- Sigurğardóttir S, Helgason A, Gulcher JR, Stefansson K, Donnelly P. 2000. The mutation rate in the human mtDNA control region. Am J Hum Genet. 66:1599-609;
- S. Matsumura and P. Forster. 2008. Generation time and effective population size in Polar Eskimos. Proc. R. Soc. B 275:1501-1508.

Sample size:

- Felsenstein, J.2005. Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? MBE 23: 691-700.
- Pluzhnikov A, Donnelly P. 1996. Optimal sequencing strategies for surveying molecular genetic diversity. Genetics 144: 1247-1262.