Topics

- Homework

- A bit more about phylogenies

- Coalescent theory

  – What is it good for?
  – How does it work?

A homework comment

- The Inland birds differ from all others by genome rearrangements

- How could that happen?

  – Genome rearrangements usually underdominant
  – Natural selection tends to weed them out

- Could be strong selection for the inversions?

A homework comment

- Easist explanation is tiny population size:

  - Genetic drift is strong
  - Mating among close relatives could produce homozygotes quickly

- Thus, speciation was probably peripatric (though could have been allopatric with a later bottleneck)
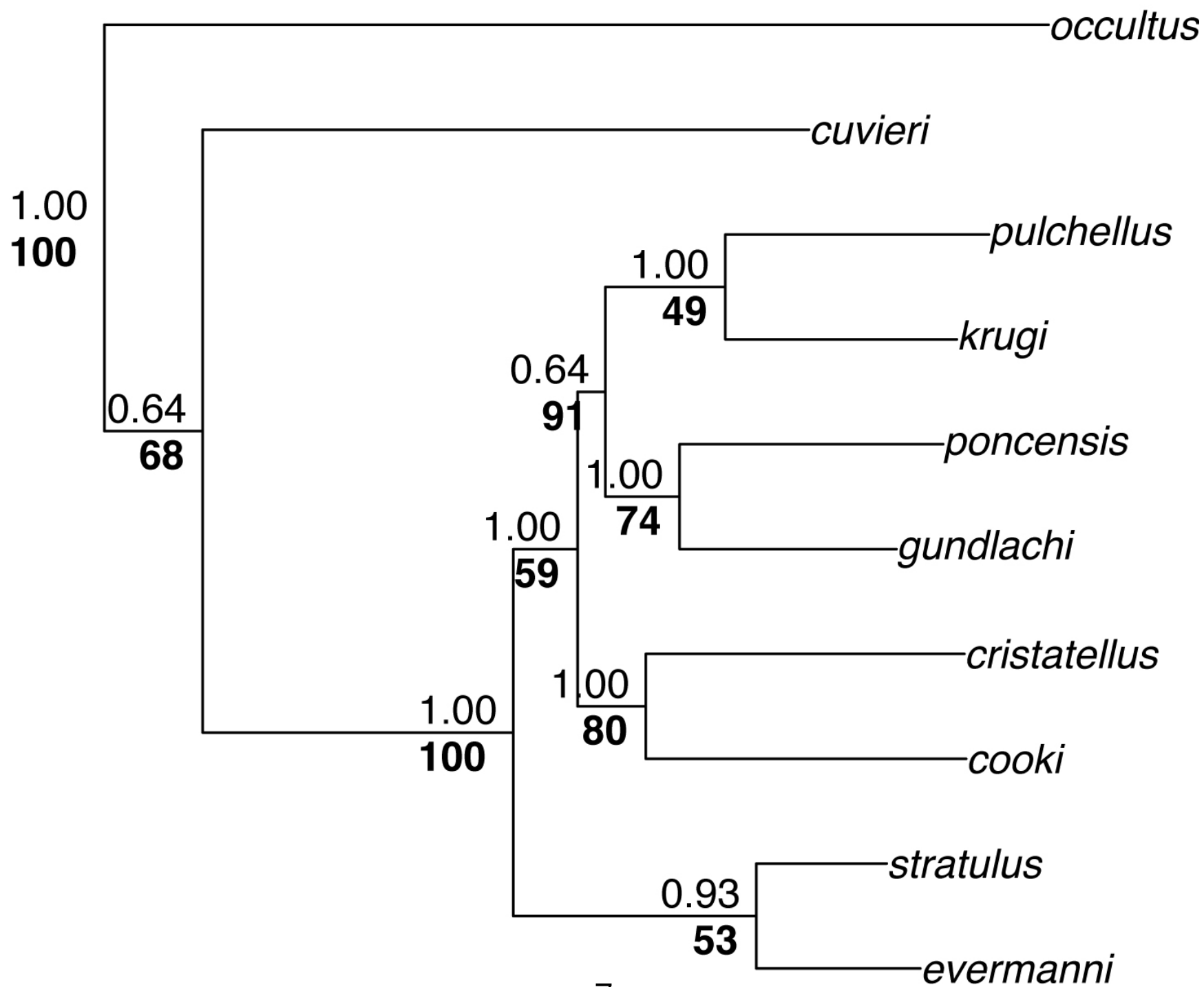
Student question

- Q: Which phylogeny methods should we be able to do by hand?

- A: Parsimony and UPGMA

  - Exam problems would be similar to HW9 problems
  - Impossible to do likelihood or Bayesian computations by hand….
  - Questions about these methods would be general/conceptual only

Phylogeny validation: likelihood

- Maximum likelihood algorithms come with built-in estimates of confidence

- Unfortunately these are only approximate for finite sized data sets

- Many researchers present bootstraps instead because they are more generally understood

Phylogeny validation: Bayesian methods

- Bayesian "cloud of trees" can be treated like a bootstrap sample

- They answer different questions:
  - Bootstrap: would a slightly different data set prefer a different tree?
  - Bayesian support: would a slightly different tree fit this data set almost as well?

- It is easier to see that these are different than to understand how to use each one appropriately!

- If "cloud" is too small, results will be overly certain

occultus

cuvieri

pulchellus

krugi

poncensis

gundlachi

cristatellus

cooki

stratulus

evermanni

1.00
100

0.64
68

1.00
49

0.64
91

1.00
74

1.00
59

1.00
80

1.00
100

0.93
53

7

Garbage in, garbage out

- No sensible tree exists when:

  - A species arose by hybridization of two other species
  - Genes have been exchanged between distantly related species
  - Different genes in the genome have different histories due to recombination and reassortment

- The programs will still run and a tree will be produced!

- Hybrids often move toward the bottom of the tree, or may cluster with one or the other parent

- Ideally we'd infer a tangled graph, but this problem is HARD

Coalescence

- Definition

- Gray whale example

- Within-population trees

- Coalescence time depends on population size

- Coalescent-based algorithms:

  - Summary statistic approaches
  - Many-tree approaches

Coalescent Theory

- Timing and pattern of common ancestry within a population reflects past population size

- It can also be perturbed by:

  - Population growth/shrinkage
  - Gene flow (migration)
  - Recombination
  - Natural selection

- If we can detect these patterns we can infer past population history

What was the long-term population size of gray whales?

Alter et al. (2007) DNA evidence for historic population size and past ecosystem impacts of gray whales. PNAS 104: 15162-15167.

What was the long-term population size of gray whales?

- How many gray whales pre-whaling?

- Whaling ship records not conclusive

- Recent slowing of the observed growth rate may suggest recovery

- Molecular data an alternative source of information
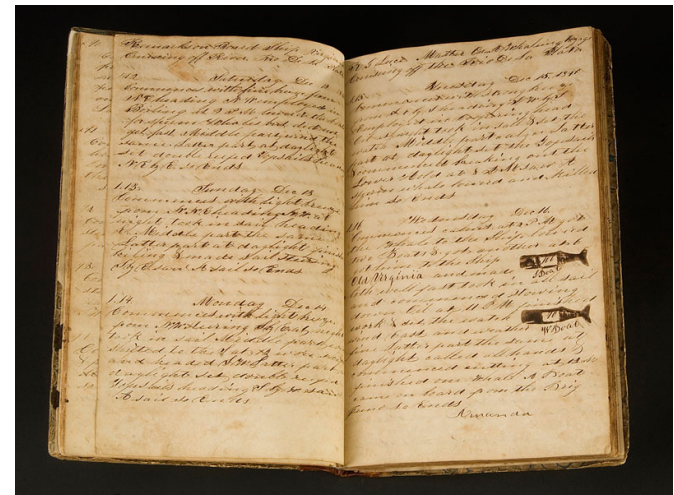
What was the long-term population size of gray whales?

- 10 loci:

  - 7 autosomal
  - 2 X-linked
  - 1 mtDNA

- Complex mutational model with rate variation among loci

- Complex population model with subdivision and copy number

- Complex demographic model relating N_{census} to N_e

	Locus	n	Estimated N
Aut	ACTA	72	162,625
	BTN	72	76,369
	CP	76	77,319
	ESO	72	272,320
	FGG	72	180,730
	LACTAL	72	44,410
	WT1	80	51,972
X	G6PD	30	2,769
	PLP	52	92,655
mtDNA	Cytb	42	107,778
	All data		96,400 (78,500-117,700)
	Current census		18,000-29,000
	Previous models		19,480-35,430

What was the long-term population size of gray whales?

- Important conservation implications

- Effect on ecosystem significant:

    - Resuspension of up to 700 million cubic meters sediment
    - (12 Yukon Rivers worth)
    - Food for 1 million sea birds

- If accepted, result suggests halving gray whale kill rate

- Broadly similar results for minke, humpback, and fin whales

Wright-Fisher population model
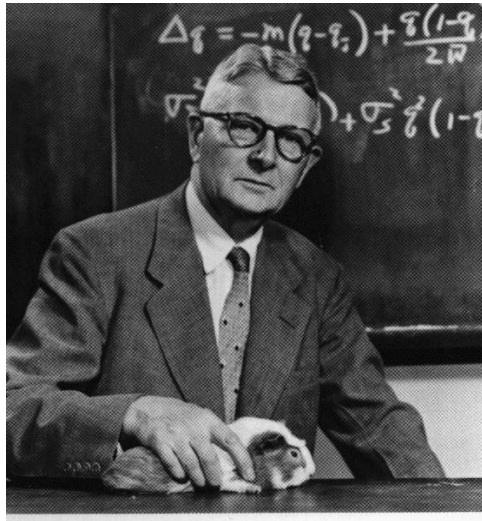
Sewall Wright showed that the probability that 2 gene copies come from the same gene copy in the preceding generation is

$$\mathrm{Prob}\ (\text{two genes share a parent}) = \frac{1}{2N}$$

Prob(having a parent)=1   Prob(having same parent)=1/(2N)
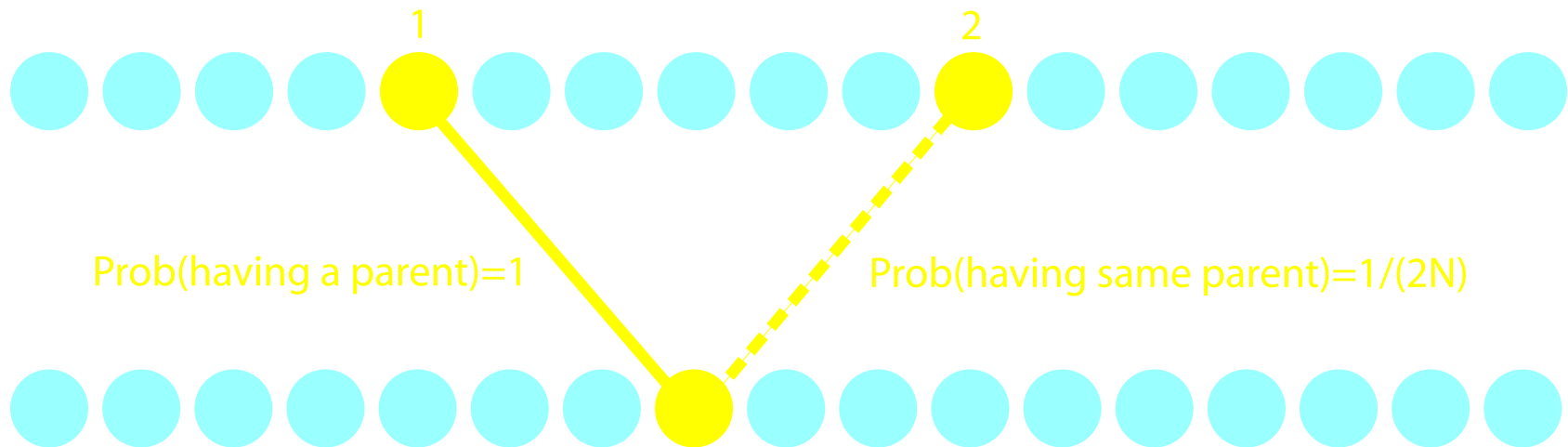
Coalescence time depends on population size

- Time back until a coalescence depends on population size N

- For k lines, the expected time (backwards) until a coalescence is $k(k-1)/2N$

- The time back to the second, third, etc. has the same type of distribution

- We can estimate N by collecting information about coalescence times

- The bigger N is, the longer the coalescence times

Coalescence time depends on population size

$$\mathbf{N} = 2270$$

$$\mathbf{N} = 12286$$

19
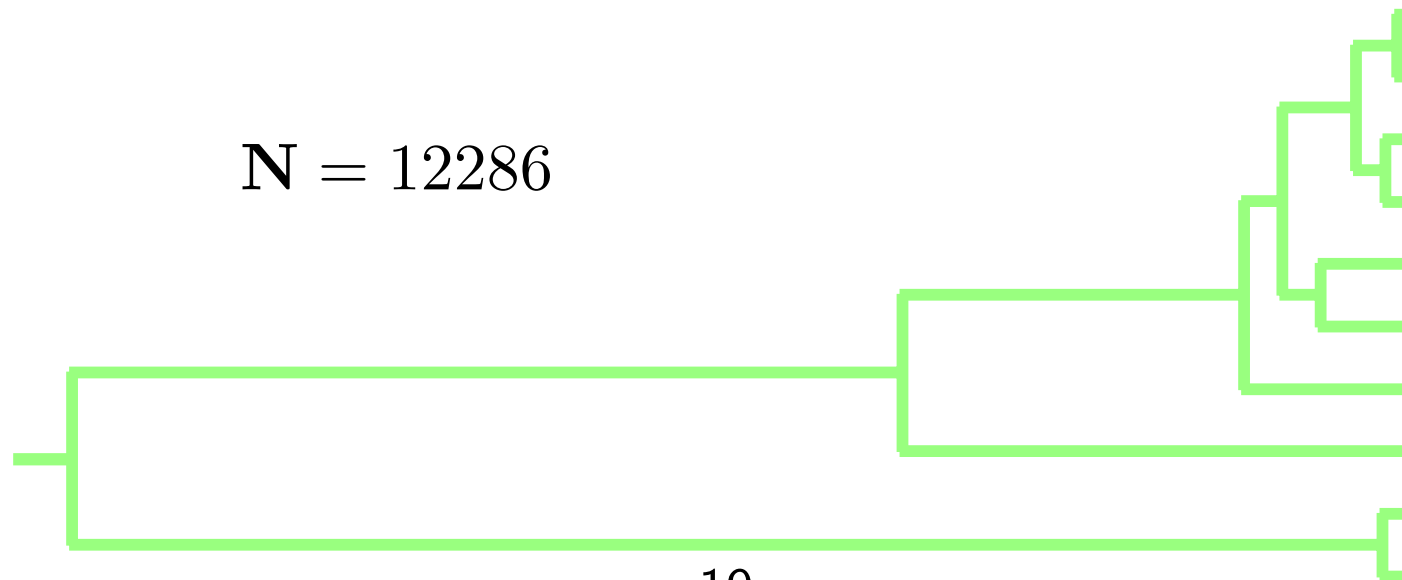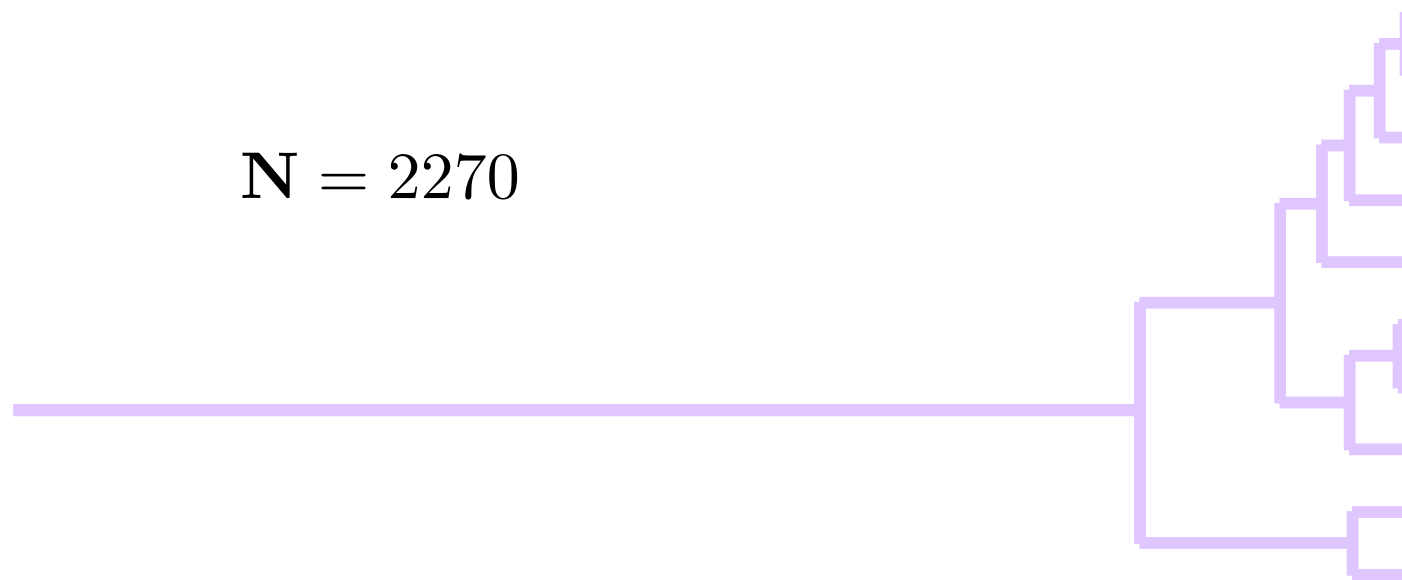
This would be great if....

- If we knew the tree, including its times, we would have a powerful estimator of population size

- Unfortunately this is difficult to infer

- Within-population variation too low for accurate phylogeny estimation

- We also have a problem with times:

  - All we observe (except with viruses or fossil DNA) is mutational differences
  - Need to know mutation rate μ to get times

The variable Θ (Theta)

- We estimate the compound parameter $4N_e\mu$ also called Θ

- One factor of 2 comes from each individual having two gene copies (so the number of gene copies is $2N$)

- The other comes from mutations accumulating on both branches of the tree, so in 1 unit of time we accumulate 2 units of mutations

The variable Θ (Theta)

- Disappointing not to get N_e directly

- If we can measure μ experimentally we can convert Θ to N_e

- Even if we can't, Θ is interesting:

  - Comparing populations with similar mutation rate
  - Expected "carrying capacity" of genetic diversity

- Examples:

  - Estimated Θ higher in Africans than other humans (expected)
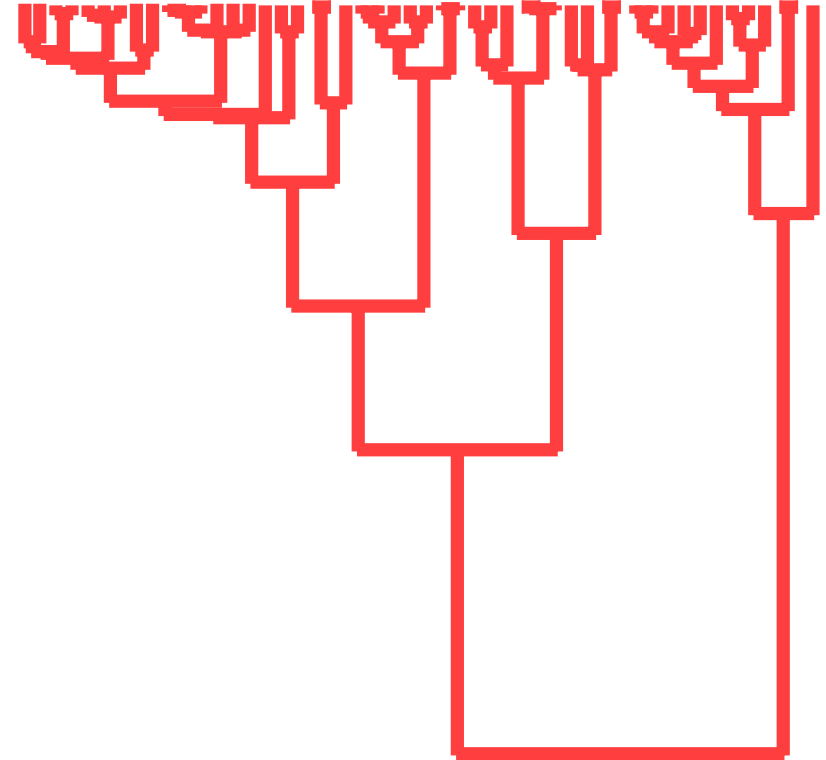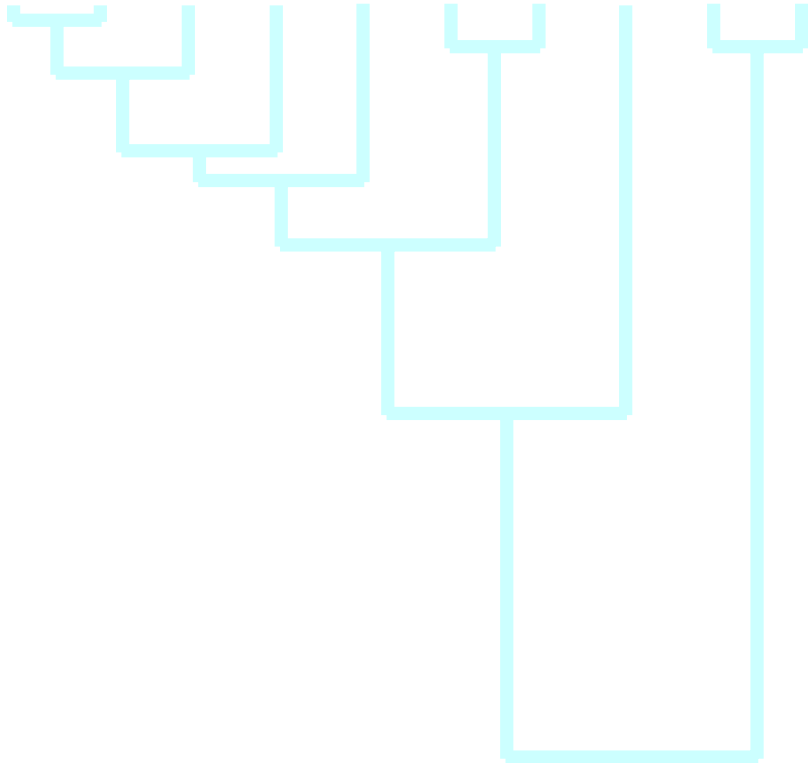 - Estimated Θ higher in chimps than humans (not expected: bottleneck in humans?)

Variability of the coalescent

A single gene can give a misleading answer:
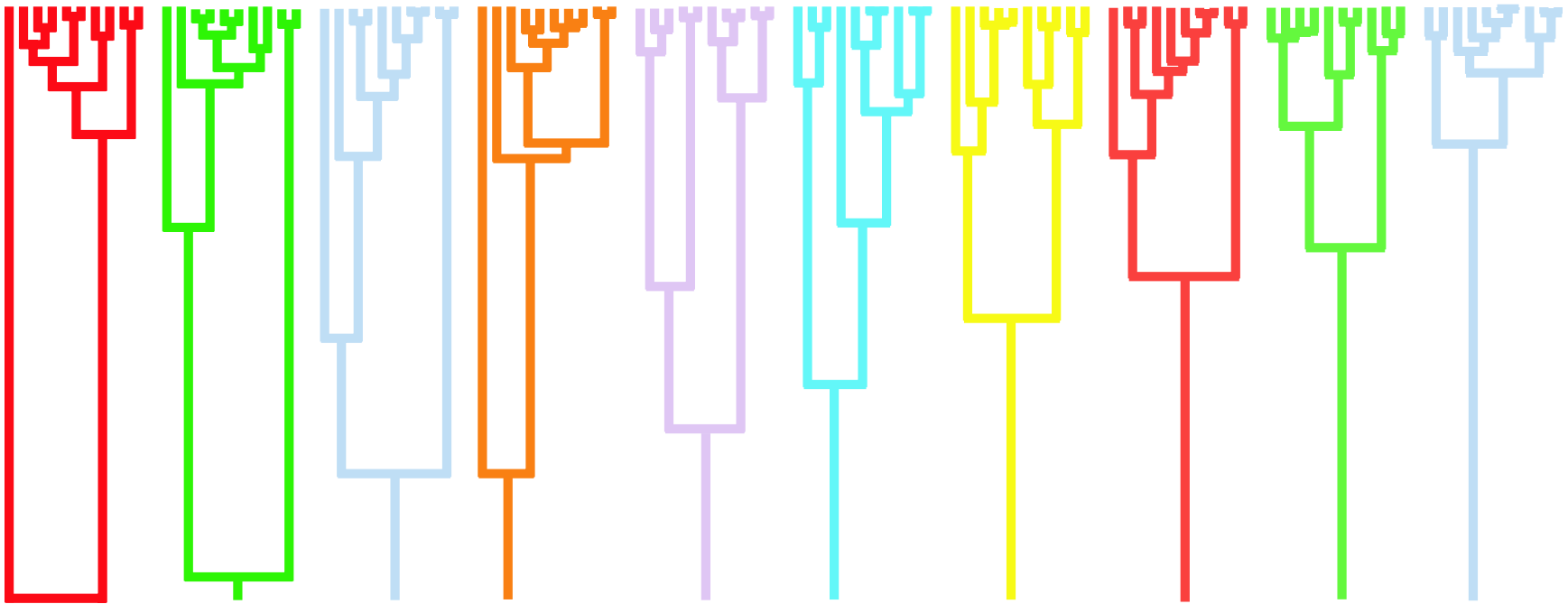
*10 coalescent trees generated with $N = 10,000$*

Summary-statistic approaches

- Summary statistics look at the bulk properties of coalescent trees.

- They often require a simplified model of mutation.

- Watterson's estimator of Θ counts variable sites

- We know how many variable sites to expect for various values of Θ, sequence length and number of sequences

- This approach discards much of the information in the data

Many-tree approaches

- My lab tries to estimate Θ by considering many possible trees

- We write sampling algorithms which visit mainly the most likely trees

- Similar to Bayesian phylogenetic algorithm

Illustration of a biased random walk

Figure generated using MCRobot program (Paul Lewis, 2001)
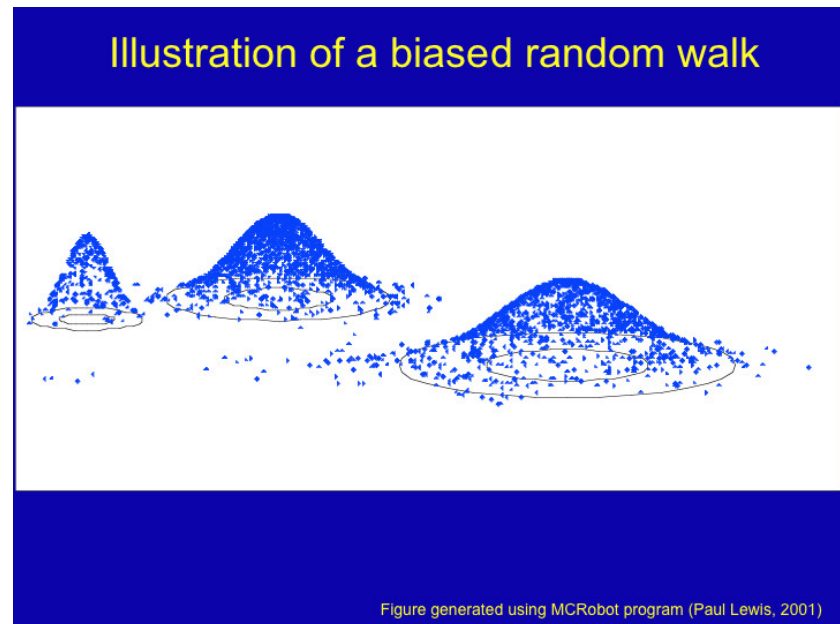
# Variants and extension of the coalescent

- Population growth/shrinkage over time

- Migration between populations

- Recombination

- Divergence of populations

- Selection (someday!)

Variable population size

- In a small population lineages coalesce quickly

- In a large population lineages coalesce slowly

This leaves a signature in the data. We can exploit this and estimate the population growth rate g jointly with the population size Θ.

Exponential population size expansion or shrinkage
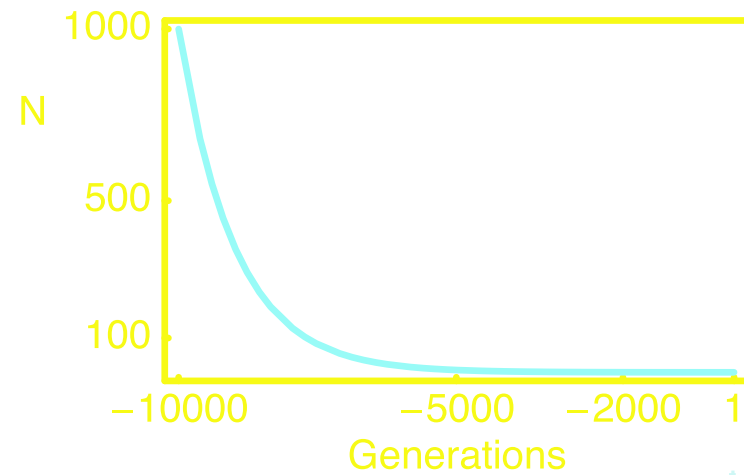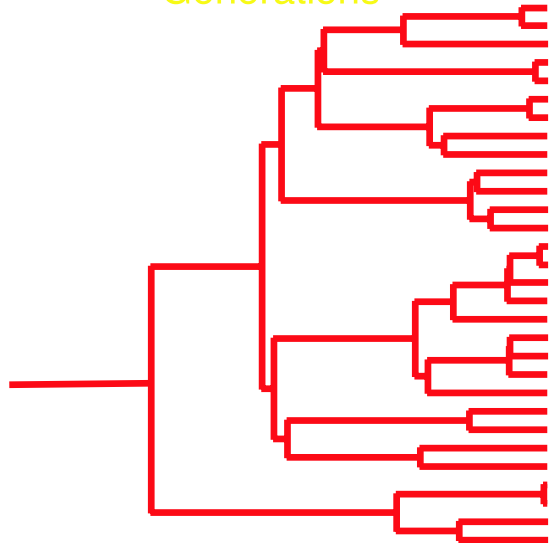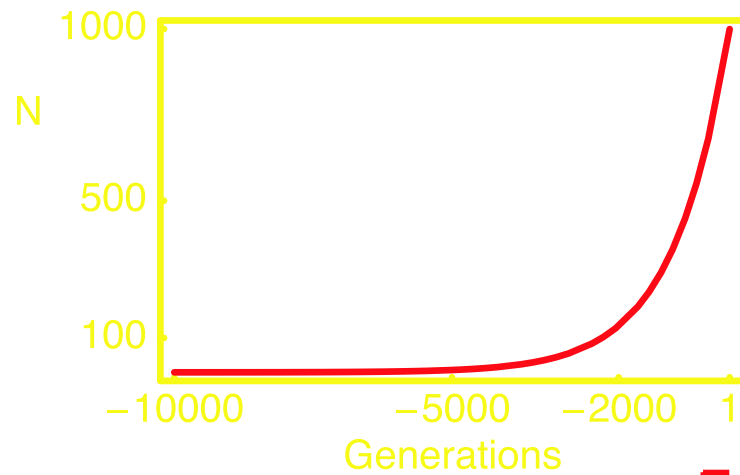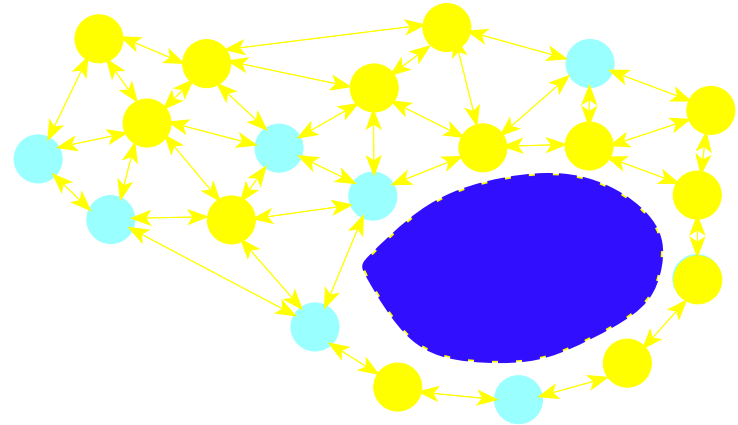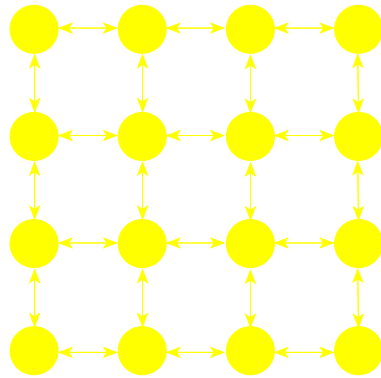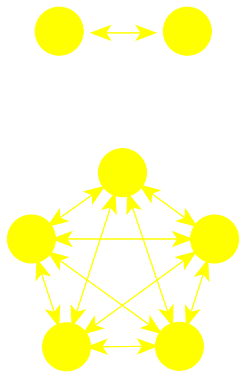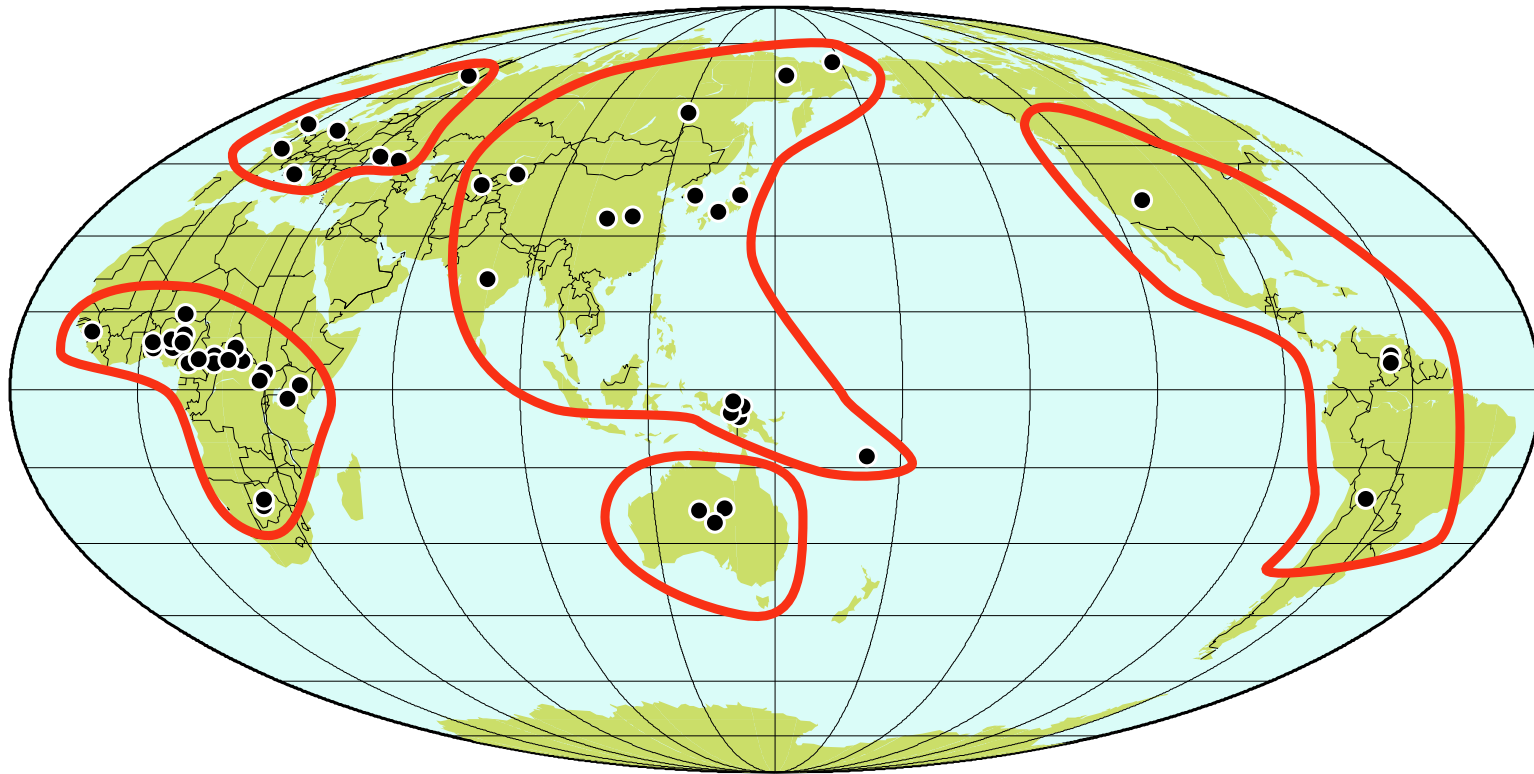
Water frog data: easier to estimate Θ than g
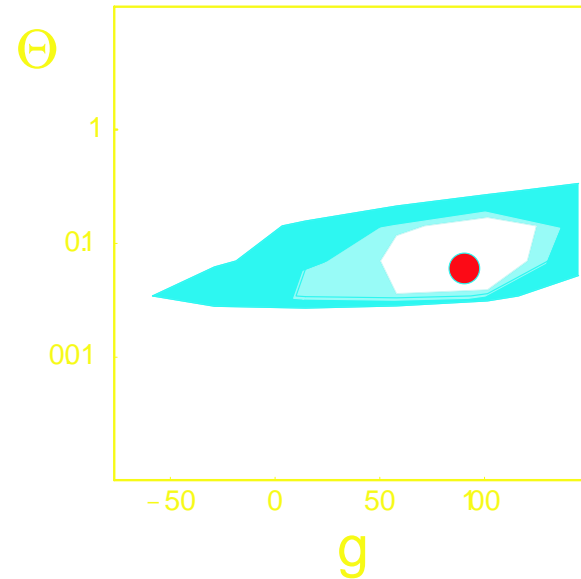
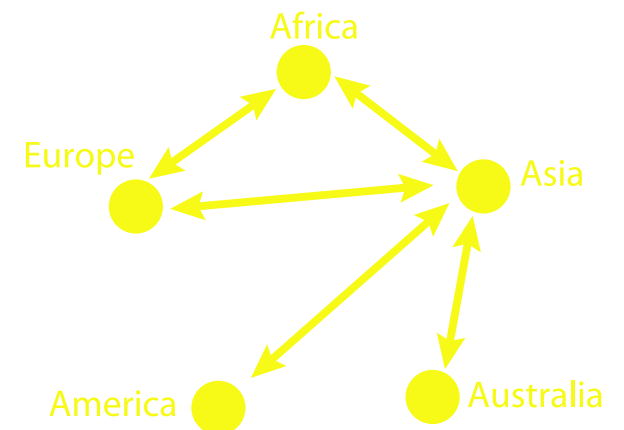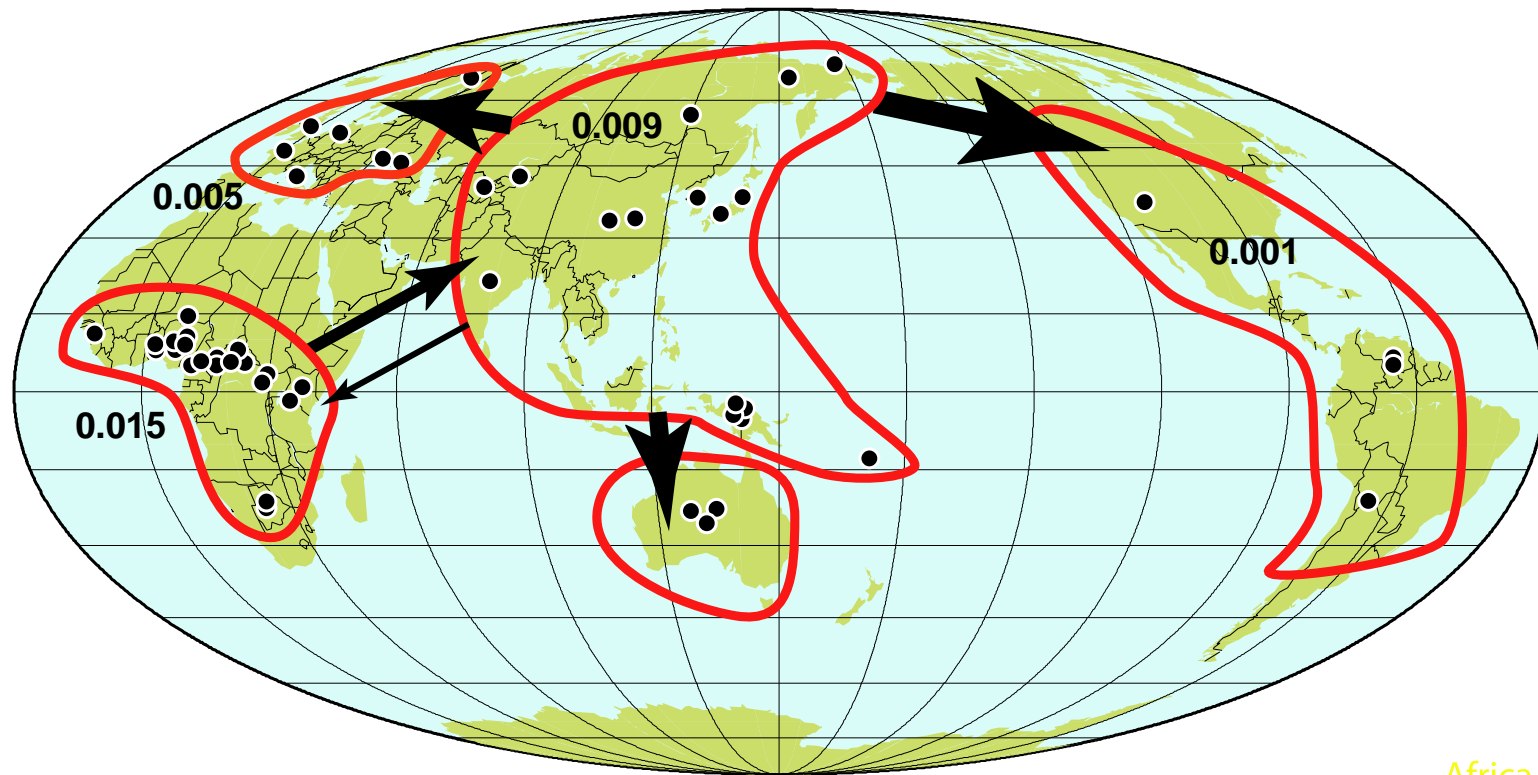| Mutation Rate	Population sizes	
	-10000 generations	Present
10^{-8}	$8,300,000$	$8,360,000$
10^{-7}	$780,000$	$836,000$
10^{-6}	$40,500$	$83,600$

Gene flow

Complete mtDNA from 5 human "populations"

A total of 53 complete mtDNA sequences ($\sim$ 16 kb):
Africa: 22, Asia: 17, Australia: 3, America: 4, Europe: 7.

Restricted model: only migration into neighbors allowed

What is the effective population size of red drum?

Turner, Wares, and Gold (2002)
Genetic effective size is three orders of magnitude smaller than adult census size in an abundant, estuarine-dependent marine fish
Genetics 162:1329-1339

Red Drum, *Sciaenops ocellatus*

What is the effective population size of red drum?

- Census population size: 3,400,000

- Effective population size: ?

- Data set:

  - 8 microsatellite loci
  - 7 populations
  - 20 individuals per population

Three approaches

1. Allele frequency fluctuation from year to year

   - Measures current population size
   - May be sensitive to short-term fluctuations

2. Coalescent estimate from *Migrate*

   - Measures long-term harmonic mean of population size
   - May reflect past bottlenecks or other long-term effects

3. Demographic models

   - Attempt to infer genetic size from census size
   - Vulnerable to errors in demographic model
   - Not well established for long-lived species with high reproductive variability

What is the effective population size of red drum?

Estimates:

Census size (N):	3,400,000
Allele frequency method (N_e):	3,516 (1,785-18,148)
Coalescent method (N_e):	1,853 (317-7,226)

The demographic model can be made consistent with these only by assuming enormous variance in reproductive success among individuals.

What is the effective population size of red drum?

- Allele frequency estimators measure current size

- Coalescent estimators measure long-term size

- Conclusion: population size and structure have been stable

What is the effective population size of red drum?

- Effective population size at least 1000 times smaller than census

- This result was highly surprising

- Red drum has the genetic liabilities of a rare species

- Turner et al. hypothesize an "estuary lottery"

- Unless the eggs are in exactly the right place, they all die

Coalescent theory–summary

- Genetic drift drives pattern of coalescences

- This embeds information about past population size in relationships among current individuals

- Various methods take advantage of this to estimate:

  - Population size, growth, shrinkage
  - Migration patterns
  - Recombination rates
  - Natural selection

Other applications of the coalescent

- Tracking expansion of an epidemic (correlation with hospital records was amazing)

- How many humans in North America pre-Columbus?

- When did the extinct Beringian bison start to decline? Was it our fault? (Required ancient DNA samples)

- Deciding if foreign medics in Libya had brought HIV virus with them (no)

One-minute responses

• Tear off a half-sheet of paper

• Write one line about the lecture:

    – Was anything unclear?
    – Did anything work particularly well?
    – What could be better?

• Leave at the back on your way out