Statistical modeling with stochastic processes

Alexandre Bouchard-Côté

Winter 2011

Plan for today

- Motivating applications and examples
  - 'Obvious' suspects: time series & spatial statistics
  - Classical problems (with a twist): density estimation, regression, classification
  - Hot topics: Natural Language Processing (NLP), Phylogenetics, Transfer/multi-task learning
- Overview of what will be covered in the course
  - Bayesian nonparametric statistics
  - Random combinatorial objects
  - Approximate inference: Monte Carlo and variational
- Background

Stochastic processes

‘A collection of random variables indexed by an arbitrary set S’

Note 1: if S if finite, then back to an ‘undergrad’ random variable, so we concentrate on S uncountable

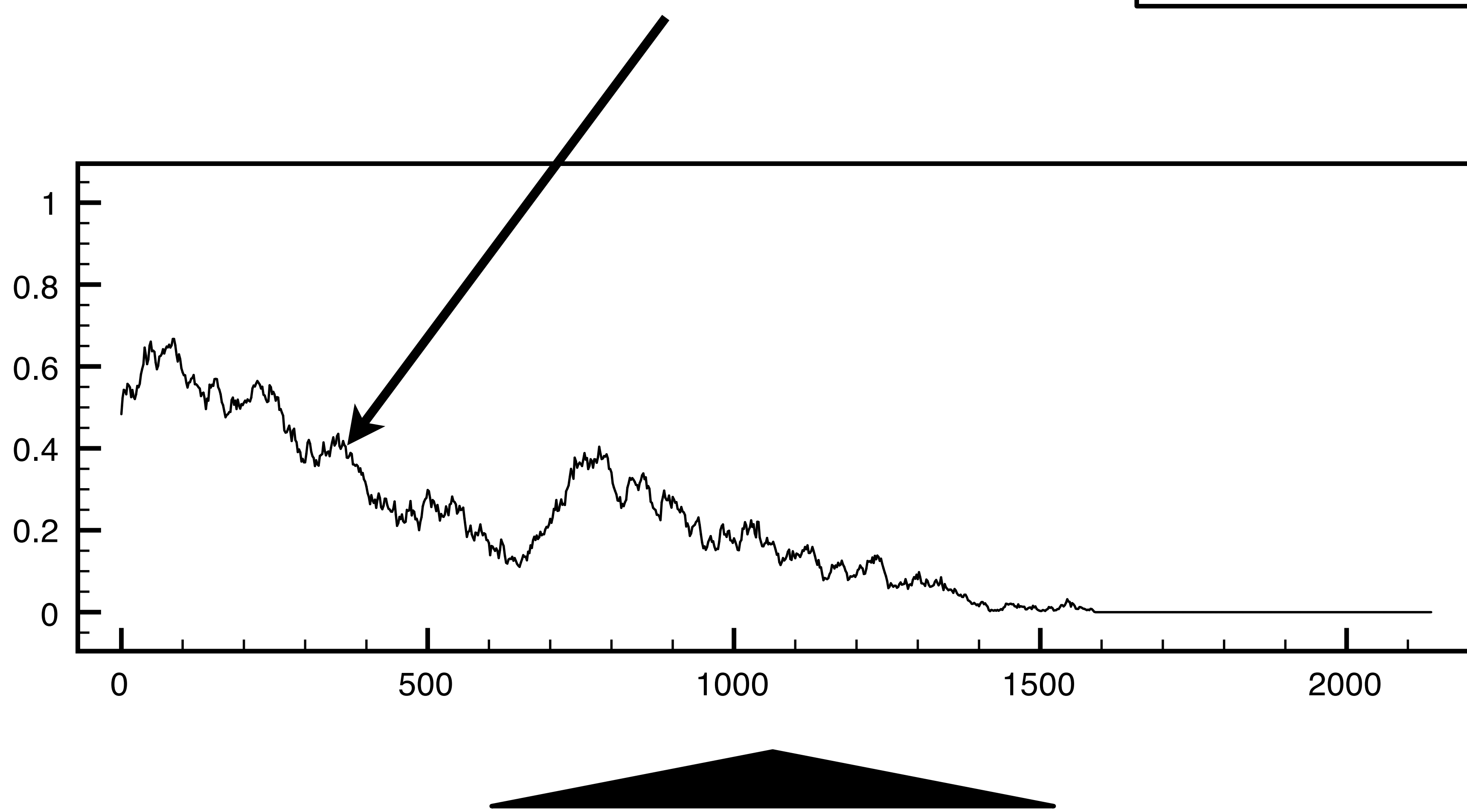
Note 2: S is not necessarily the real line

Example: distribution over functions

Samples: functions $f: \mathbf{R}^2 \rightarrow \mathbf{R}$

$$(s,\ Y_s(\omega))$$

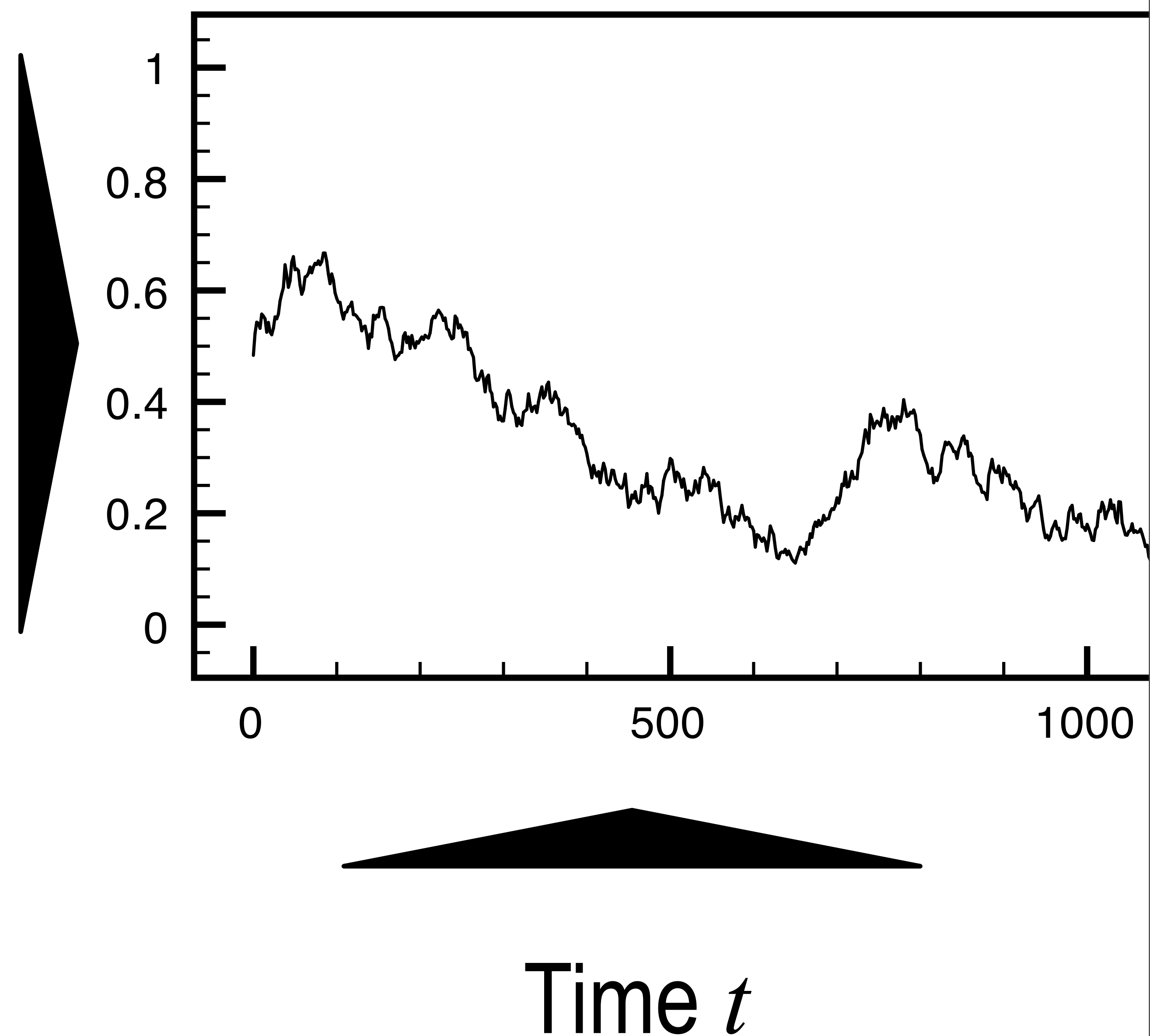
$$\boxed{Y_s(\omega) = f(s)}$$



$$S = \mathbf{R}$$

‘Obvious’ suspects

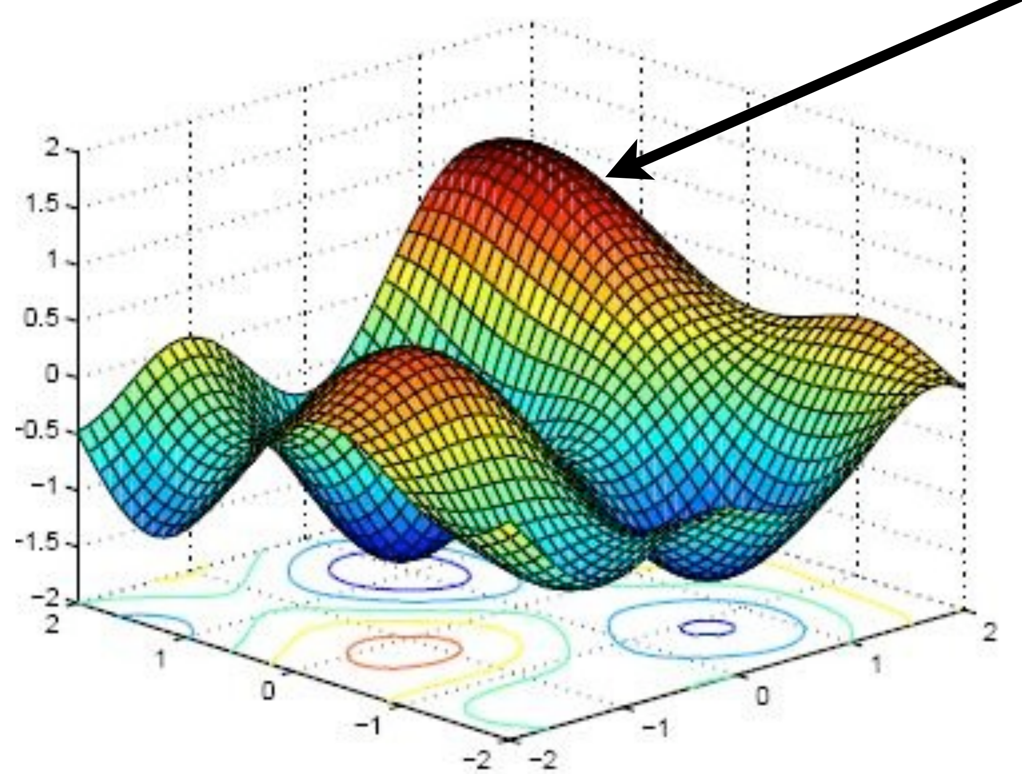
- Time series:
  - Economic/financial indicators
  - *Frequency of the population having a certain genetic mutation*
  - Global weather/climate observations



Example: distribution over functions

Samples: functions $f : \mathbf{R}^2 \to \mathbf{R}$

$(s,\ Y_s(\omega)\ )$

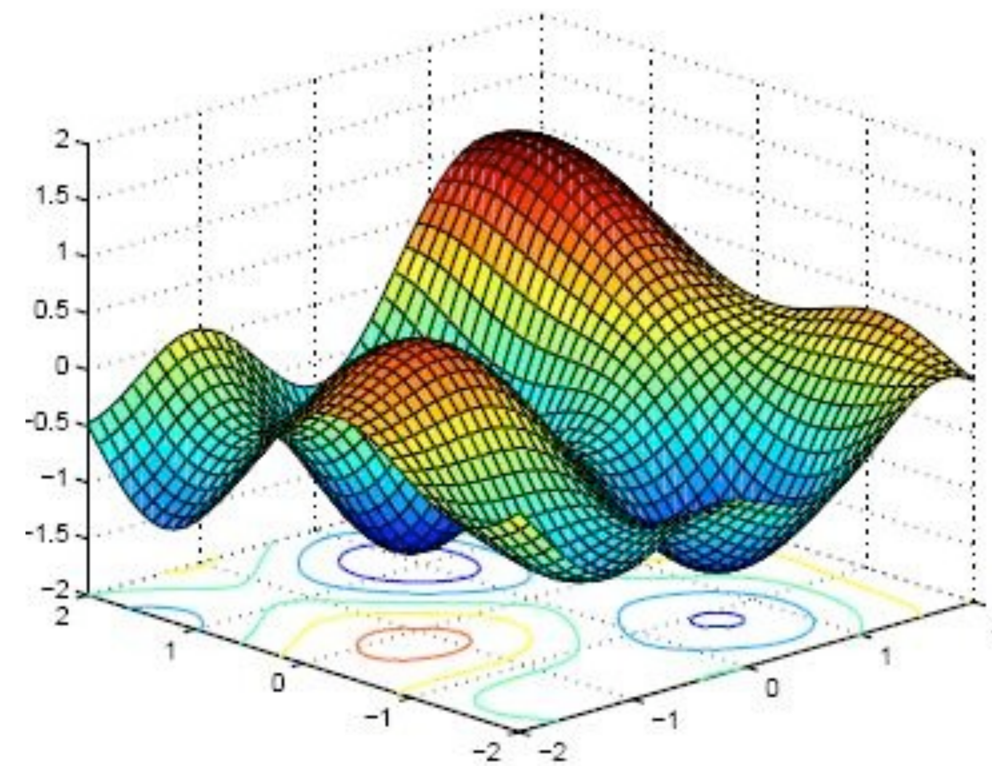


$$\boxed{Y_s(\omega) = f(s)}$$

$S = \mathbf{R}^2$

‘Obvious’ suspects

- Spatial statistics:
  - Epidemic outbreak intensity
  - Ecological measurements
  - Intensity of the cosmic background radiation

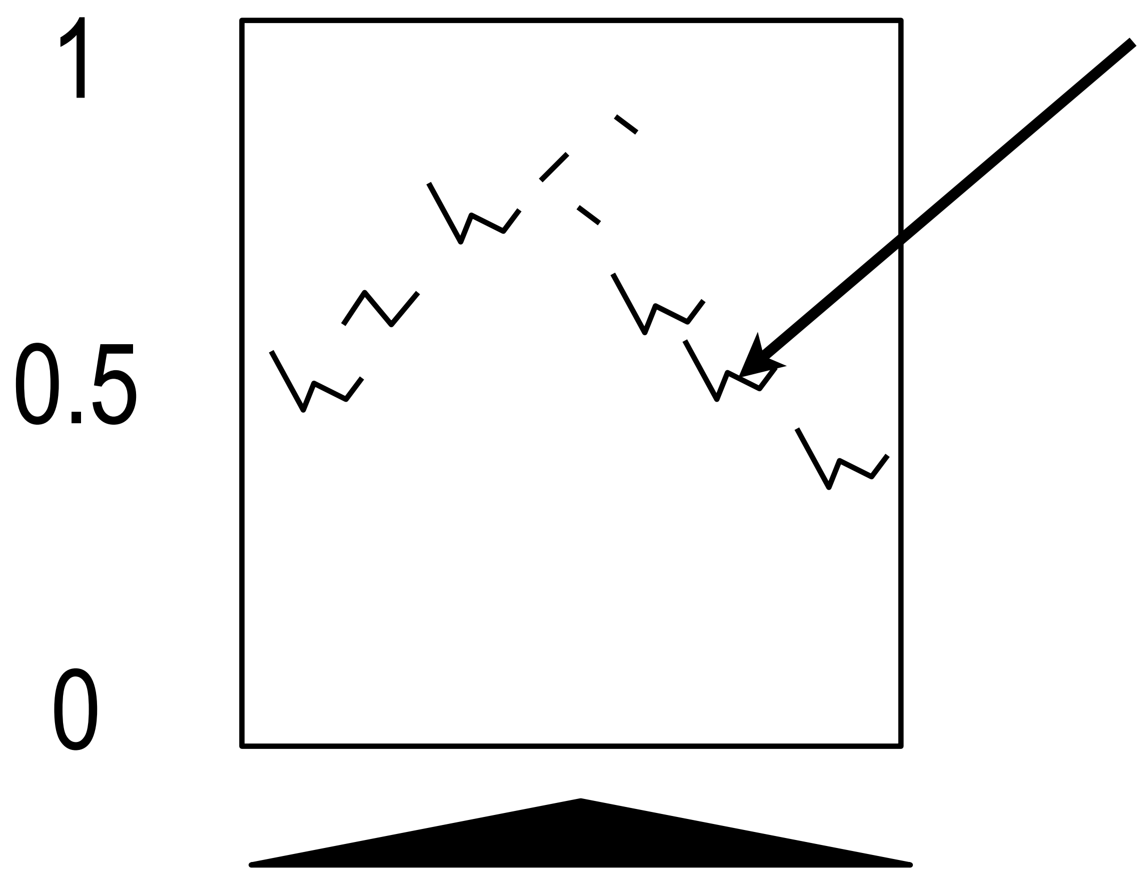


Location *(x,y)*

Example: distribution over *distributions*

Samples: distributions $\lambda : \mathcal{F} \rightarrow [0,1]$

$(s,\ Y_s(\omega)\ )$



$$\boxed{Y_s(\omega) = \lambda(s)}$$

$S = \mathcal{F}$, a sigma-algebra (the set of events for λ)

(No topology on this axis this time...)

Why would we need distributions over distributions?

De Finetti theorem: a compelling motivation for priors on parameters...

Suppose: we agree that if our data x_i are reorder, it doesn't matter (exchangeability), e.g.

$$(x_1, x_2, x_3, \ldots) \overset{\mathrm{d}}{=} (x_3, x_1, x_2, \ldots)$$

Then: there exists a random variable θ and distributions F_θ such that:

$$x_i | \theta \sim F_\theta$$

De Finetti theorem

In other words: if you assert exchangeability, it is reasonable to act as if there is:
- an underlying parameters,
- a prior on that parameter, and
- the data is generated independently conditionally on that parameter

Note: the theorem would not be true if we limited ourselves to random variables θ with domain $\mathbf{R}^n$
In particular, we need to allow to have distribution-valued random variables θ, hence we need priors over distributions!

Consequence

Stochastic processes can sneak out in any inference problem, not only in the standard stochastic process application ‘niches’ (i.e. time series and spatial statistics)

Example: density estimation

Input/observations: Samples of UBC students' heights x_i

Examples of inferential problems:
What is the mean height of the UBC student population?
What is the most 'atypical' height among the samples x_i?
...

Method 1 (Normal density estimation): Find a normal density ϕ_{μ,σ^2} that best fits the data

Bayesian normal density estimation

Input/observations: Samples of UBC students' heights x_i

Bayesian way: Treat the unknown quantity ϕ_{μ,σ^2} as random. Equivalently: treat the parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ as random.

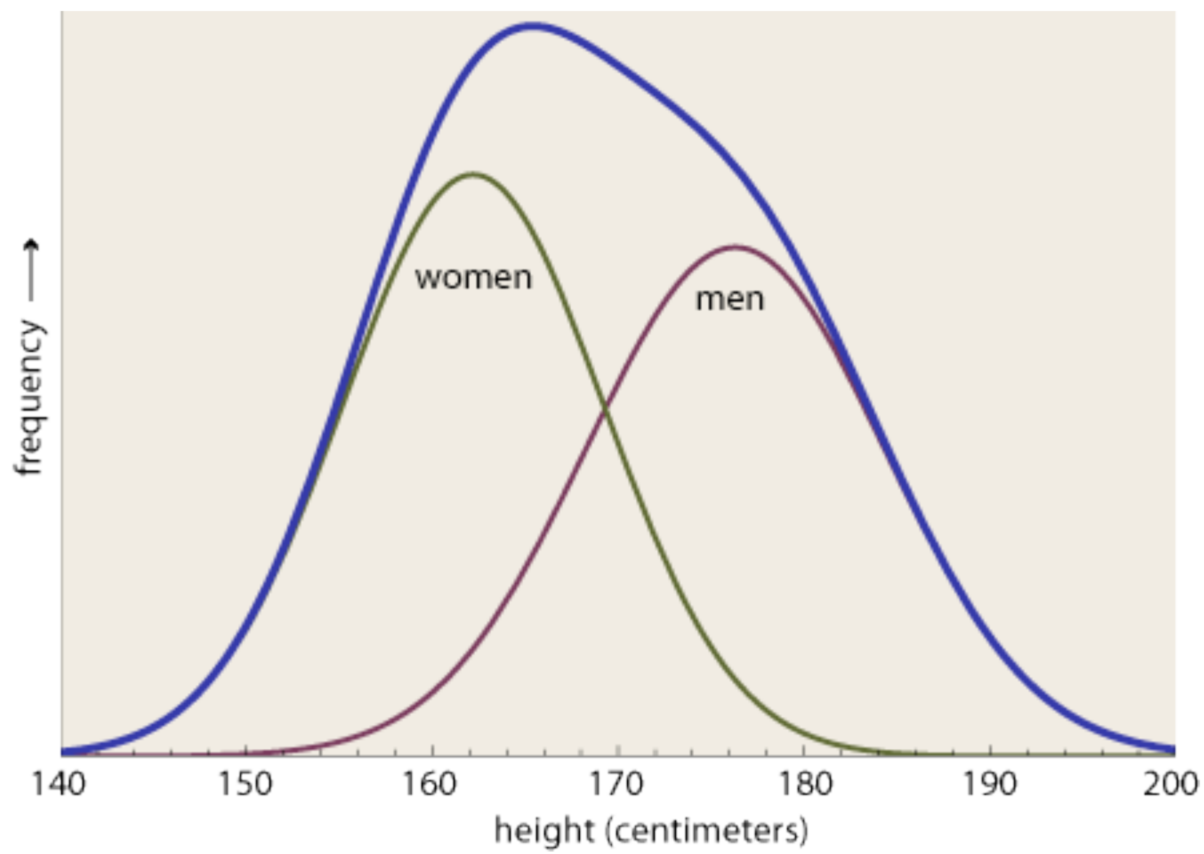
Output: Posterior over densities / the parameters of a normal distribution

Details of the model: Not critical for now, but will be needed later...

Bayesian normal density estimation

Limitation: fails to model that men and women have different height distributions!

Solution: use a *mixture model* with two mixture components, each one assumed to be normal



Density estimation of normal mixtures

But we did not recorded the male/female information when we collected the heights!

Expensive fix: Do the survey again, collecting the male/female information

Cheaper fix: Let the model guess, for each datapoint, from which *cluster* (group, mixture component) it comes from.

Method 2 (Mixture of two normal distributions)

Bayesian way: Treat the parameters of each cluster as random: $\mu_c \in \mathrm{R}$ and $\sigma_c^2 > 0$, $c \in \{1,2\}$

The variables $z_i \in \{1,2\}$ indicate which cluster observation i belongs to (cluster membership indicator). Treat them as random as well.

The parameters π_c are priors over the cluster indicators (fraction of male vs. female at UBC). Treat them as random.

Closely related to: *unsupervised learning*

Method 2 (Mixture of two normal distributions)

There are still limitations to this model:

- Height distribution also depends on the age of the student
- Height distribution also depends on the ethnicity of the student
...

Idea: Use more than two mixture components!

Using more mixture components

- Should we make the number of clusters as large as possibles?

Using more mixture components

- Should we make the number of clusters as large as possibles?

- How many clusters should we use?
  - Methods you are familiar with: using cross validation, AIC, BIC, etc.
  - Another route: non parametric Bayesian priors

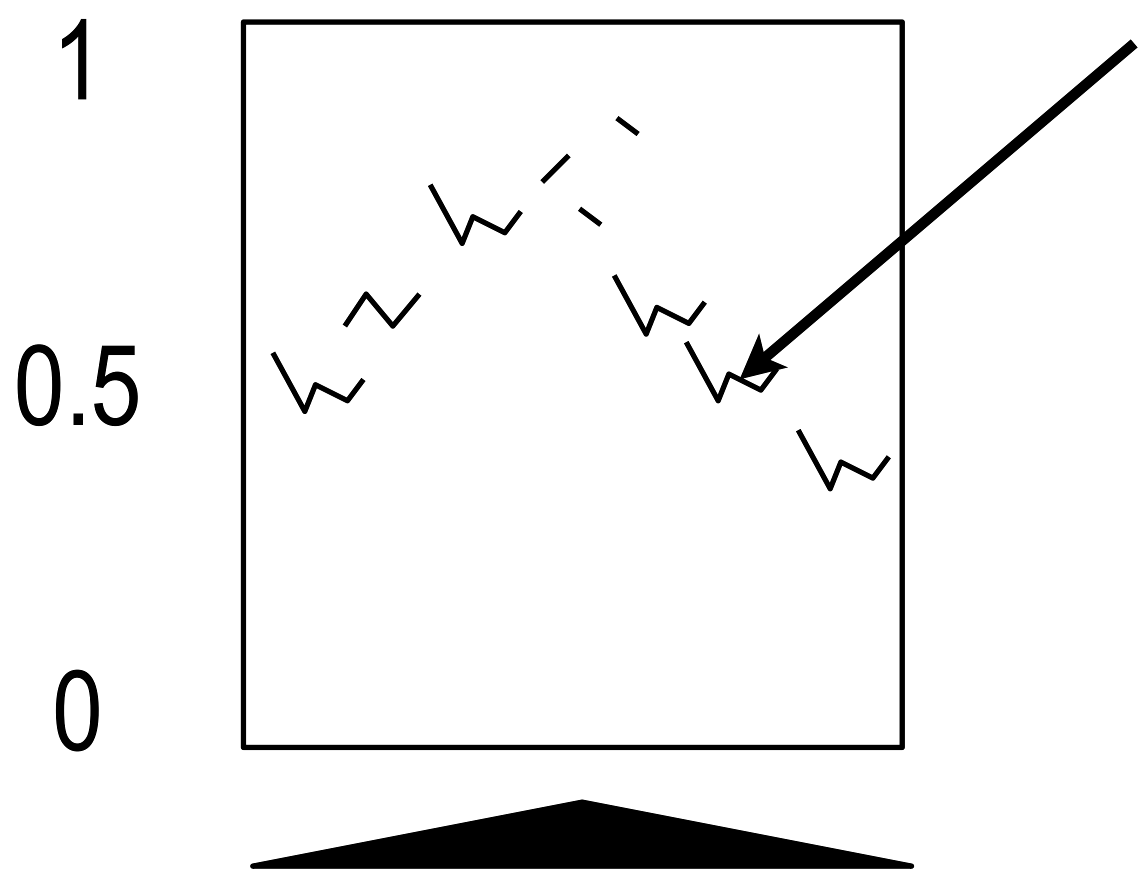
- Rough idea of non parametric Bayesian statistics
  - Prior allowing a countably infinite number of clusters while giving protection against over-fitting
  - Claim: this prior takes the form of a distribution over distributions...

Example: distribution over *distributions*

Samples: distributions $\lambda : \sigma(\Theta) \rightarrow [0,1]$

$(s,\ Y_s(\omega)\ )$

$$\boxed{Y_s(\omega) = \lambda(s)}$$



$$S = \sigma(\Theta)$$

(No topology on this axis this time...)

Applications in Natural Language Processing

Language models

Shannon's game: guess the next word...

I have lived in San ______

I am not going to go ______

there or their?

Application: finding which sentence is more likely

Example: Speech recognition

Language models: first approach

Fix a certain prefix length, and estimate one categorical distribution for each prefix from a text dataset (n-gram)

Distribution over what follows after the prefix
Fix ___

Guess	Pr
a	1.0

Distribution over what follows after the prefix
a ___

Guess	Pr
certain	0.5
text	0.5

...

Problem with the maximum likelihood estimator?

Language models: second approach

Prior for prefix 1	Prior for prefix 2	...

Prior for prefix 1

Distribution over what follows after the prefix Fix ___

Guess	Pr
a	0.92
...	...
...	...

Prior for prefix 2

Distribution over what follows after the prefix a ___

Guess	Pr
certain	0.46
text	0.46
...	...

...

...

Some prefixes are rare. Is that a problem?

Language models: third approach

Hyper-prior over words---not specific to a prefix

Distribution over words in text dataset

Guess	Pr
the	0.04
a	0.02
...	...

Prior for prefix 1

Distribution over what follows after the prefix
Fix ___

Guess	Pr
a	0.92
...	...
...	...

Prior for prefix 2

Distribution over what follows after the prefix
a ___

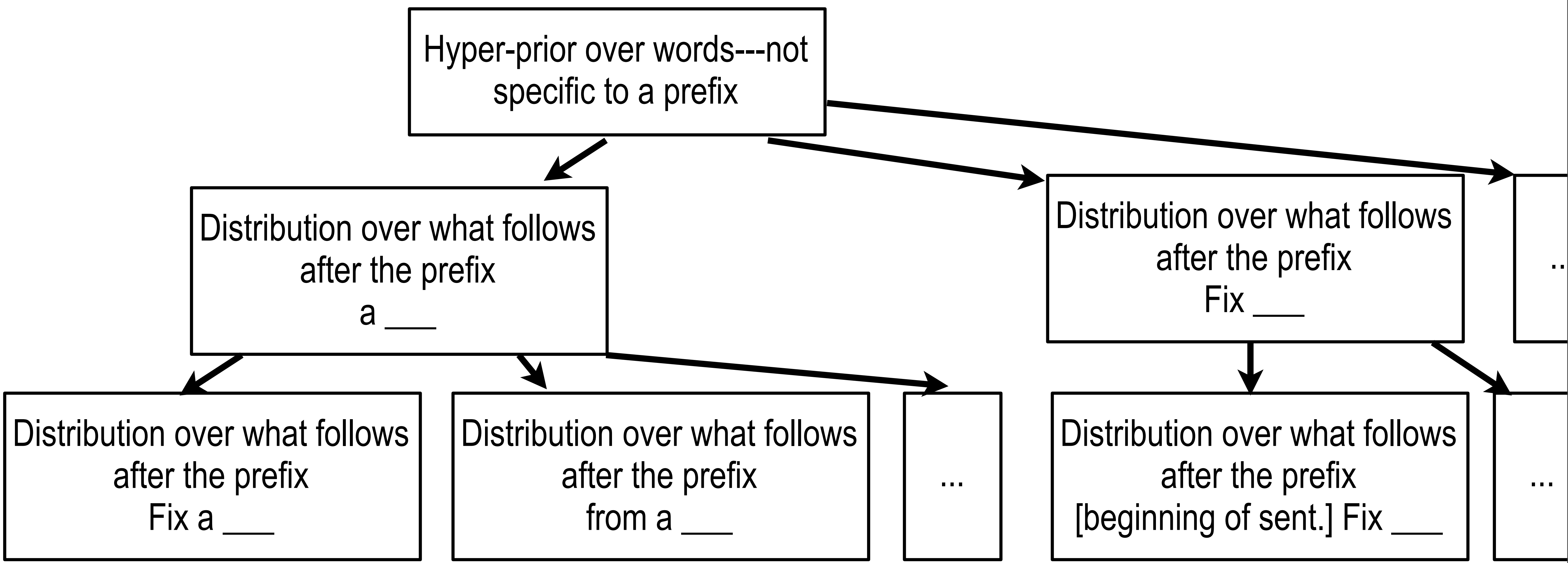
Guess	Pr
certain	0.46
text	0.46
...	...

...

...

Language models: fourth and fifth approaches

Why stop at prefixes of length 1?



Why stop at prefixes of a bounded length?

Machine translation

Ultimate goal: Pairs of Chinese-English sentences

((to build 500 gas stations, 建立500个加油站), ...)

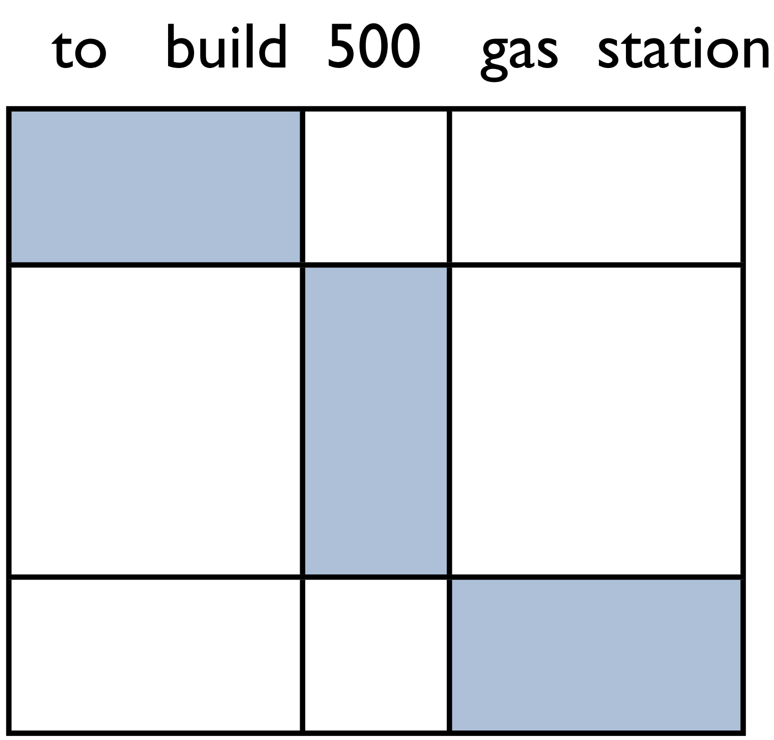
Inferential problems: Given a new Chinese sentence, translate it to English

Machine translation: Intermediate goal

Input/observations: Pairs of Chinese-English sentences

((to build 500 gas stations, 建立500个加油站), ...)

Inferential problems: Segment and align

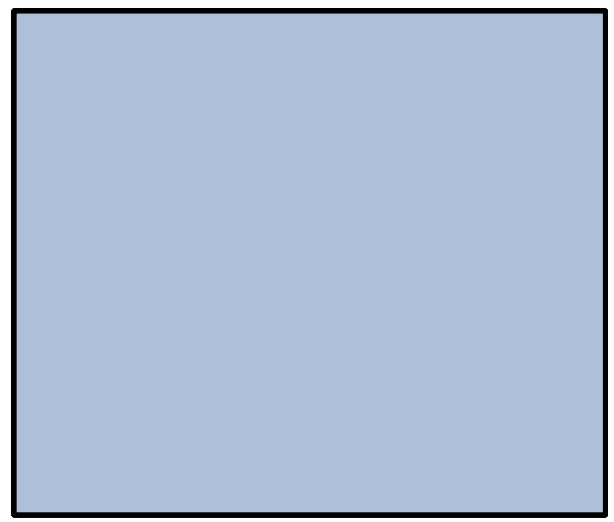


	Pinyin	Gloss
建立	jian4 li4	*build*
500		
个	ge4	*[measure word]*
加油站	jia1 you2 zhan4	*gas station*

Slide from John DeNero

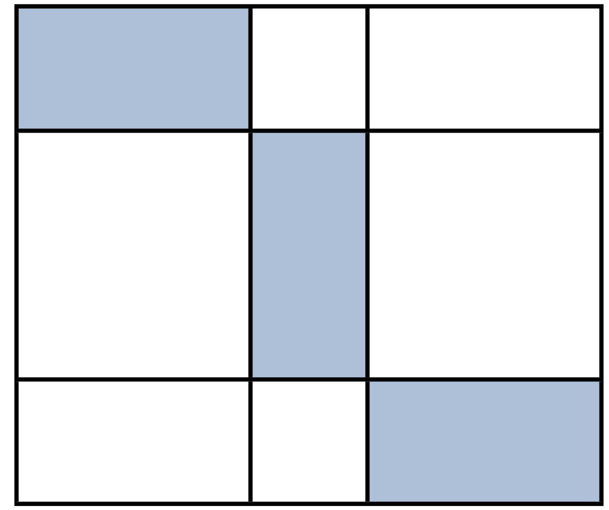
Degeneracy of previous maximum likelihood estimators

Maximum likelihood



	Pinyin	Gloss
建立	jian4 li4	*build*
500		
个	ge4	*[measure word]*
加油站	jia1 you2 zhan4	*gas station*

Non parametric Bayesian prior



	Pinyin	Gloss
建立	jian4 li4	*build*
500		
个	ge4	*[measure word]*
加油站	jia1 you2 zhan4	*gas station*

Parts of speech

Shannon's game: guess the next word...

That's something I _____.

Part of speech: a category of words defined by how the word behave in the sentence.

Examples: verbs, nouns, adjectives, adverbs, etc.

Classification problem: predicting the part-of-speech

Input/observations: Annotated sentences

NOUN	VERB	ADJ	NOUN
Alex	likes	red	apples

VERB	ADV	VERB	ADV
Talk	faster,	eat	slower

Inferential problem 1: find the part of speech of the last word in a sentence

NOUN	VERB	ADJ	?????
Alex	likes	big	houses

Predicting the parts-of-speech: cues

NOUN	VERB	ADJ	?????
Alex	likes	big	houses

What is the part-of-speech (POS) of ‘houses’?

Two cues: What POSs can follow an adjective (ADJ)?
ADJ, NOUN, but probably not VERB

What POSs can be assigned to houses?
VERB, NOUN, but probably not ADJ

Method: Hidden Markov models...

Sequential prediction

Input/observations: Annotated sentences

NOUN	VERB	ADJ	NOUN
Alex	likes	red	apples

VERB	ADV	VERB	ADV
Talk	faster,	eat	slower

Inferential problem 2: find all the parts of speech of a new sentence

???	???	???	???
Alex	likes	big	houses

Sequential clustering

Input/observations: ~~Annotated~~ sentences

NOUN	VERB	ADJ	NOUN
Alex	likes	red	apples

VERB	ADV	VERB	ADV
Talk	faster,	eat	slower

Inferential problem 3: find all the ‘parts of speech’ (clusters) of a new sentence

???	???	???	???
Alex	likes	big	houses

Sequential clustering: how many clusters?

Can use methods similar to the earlier density estimation example

Twist:
Earlier: A prior over countably infinite distributions vs.
Now: A prior over countably infinite transition matrices

Also useful when supervision (annotation) is available, each class (POS) is expressed as its own infinite mixture (*state splitting*)

Choice models

Input: Number of times people chose the row object over the column object.

	Phone 1	Phone 2	Phone 3
Phone 1	-	2	7
Phone 2	6	-	7
Phone 3	1	1	-

7 people chose Phone 1 over Phone 3

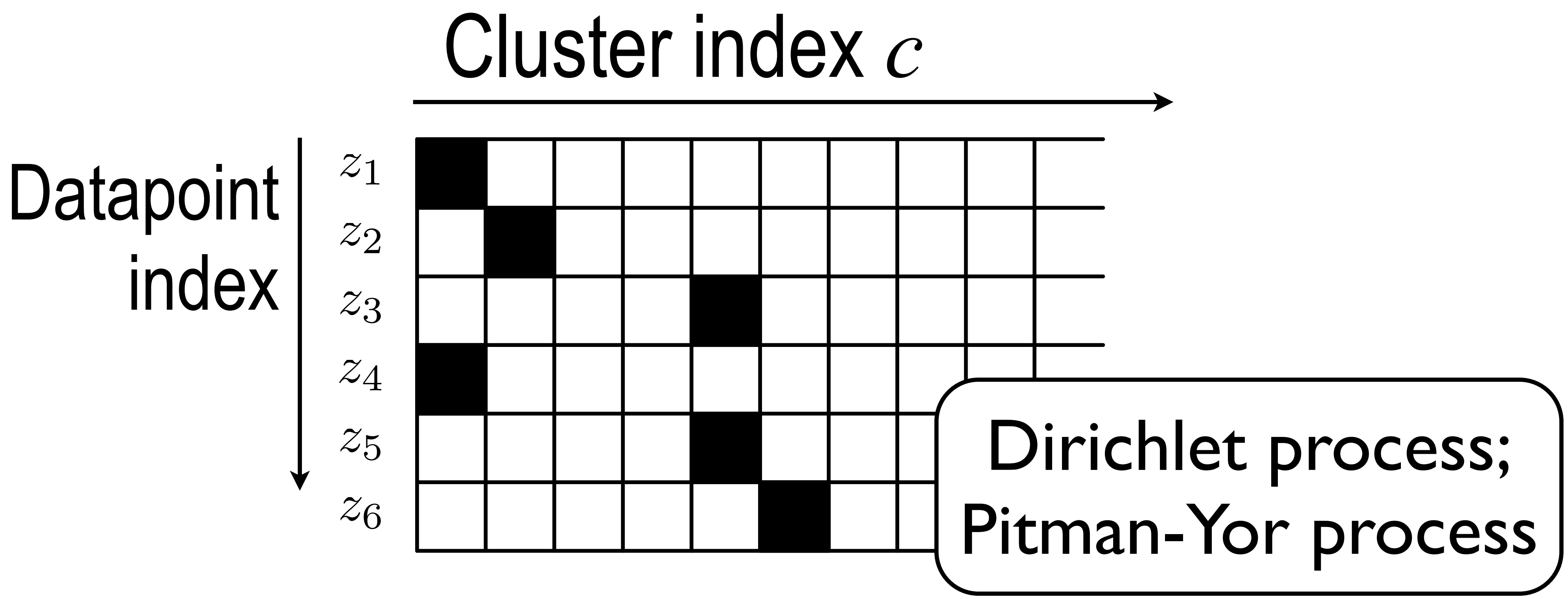
Desired output: latent features governing these choices

	Phone	Camera	Internet	Flip-phone	Cheap
Phone 1	✓	✓	✓		
Phone 2	✓	✓			✓
Phone 3	✓		✓	✓	

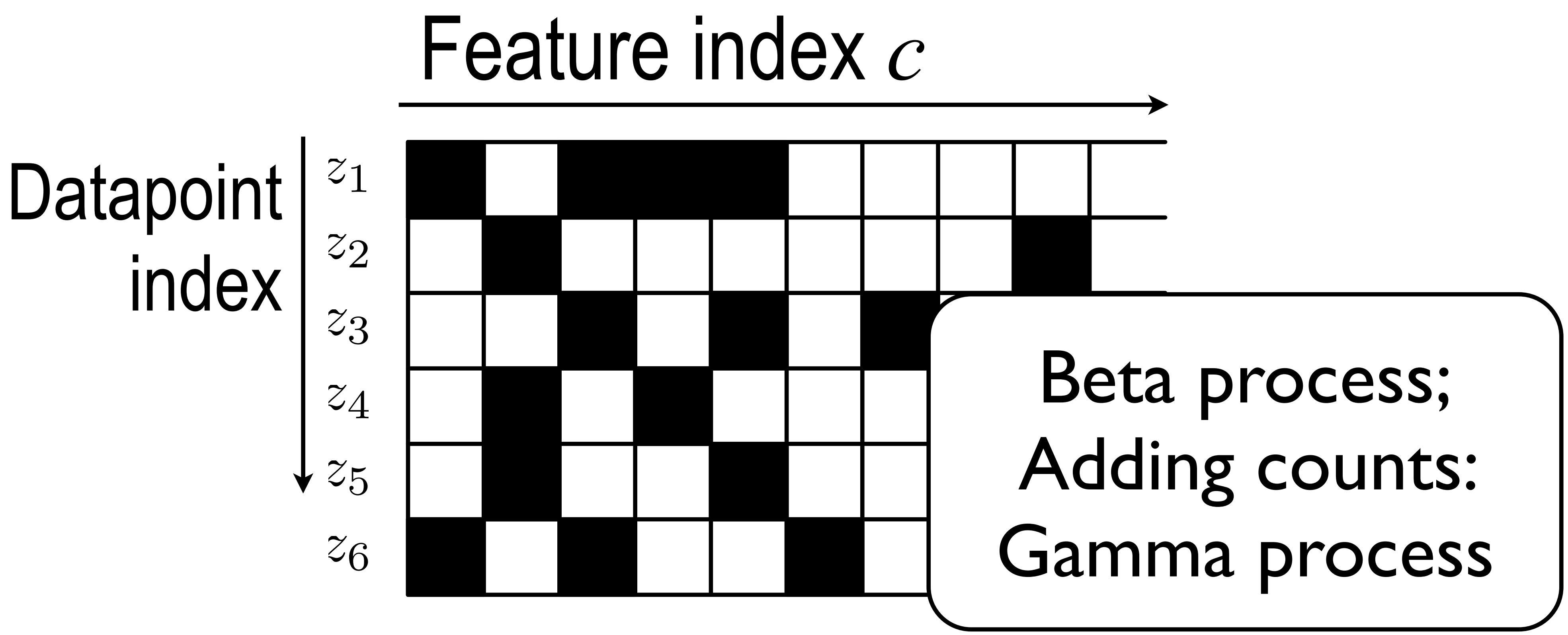
Slide from Kurt Miller

Different type of prior needed

Mixture indicator priors:



Feature indicator priors:



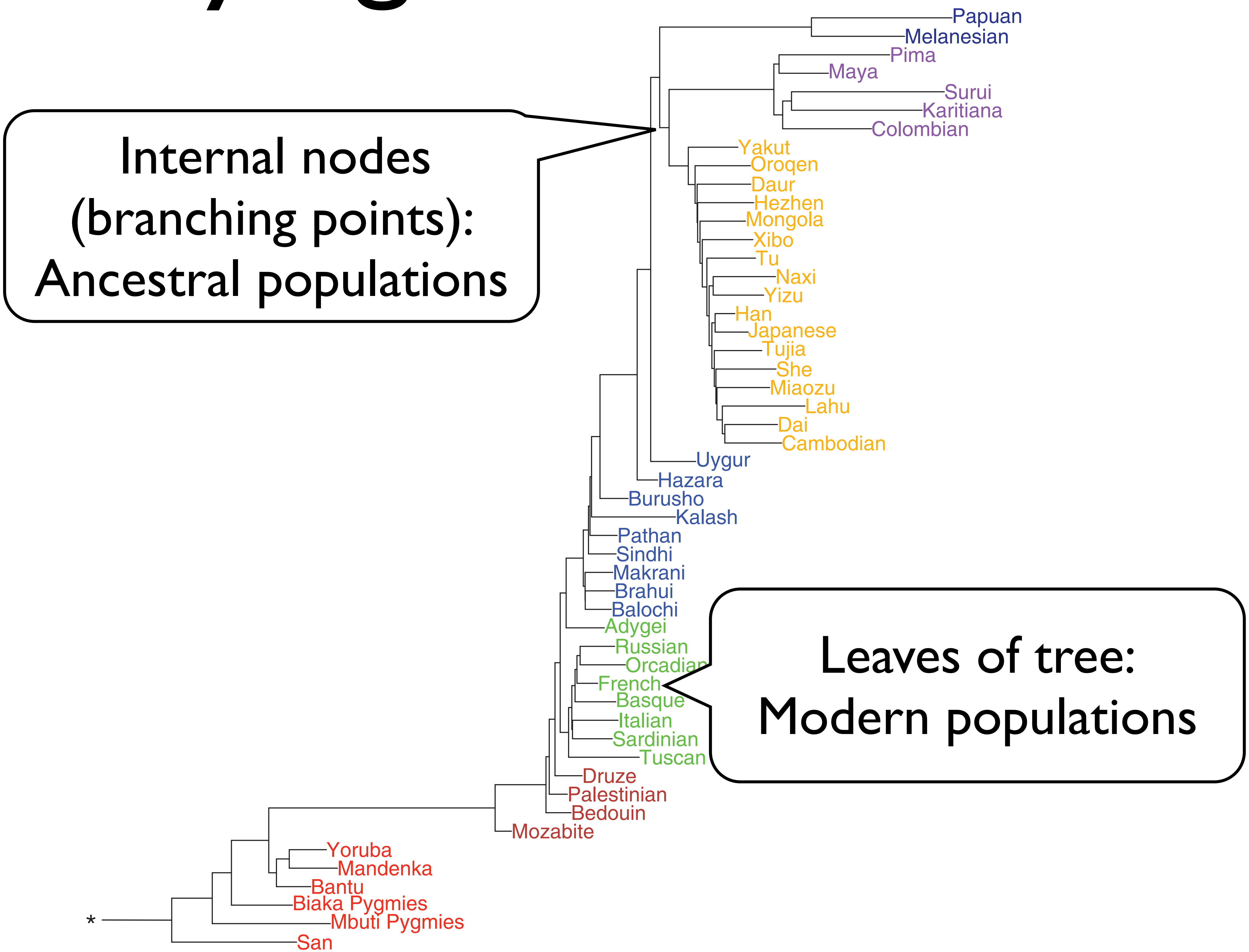
Applications in Phylogenetic Inference

Non-Bayesian application: phylogenetic inference

Scientific applications: biology, anthropology, linguistics

Engineering applications: domain adaptation, multi-task learning

Phylogenetic tree

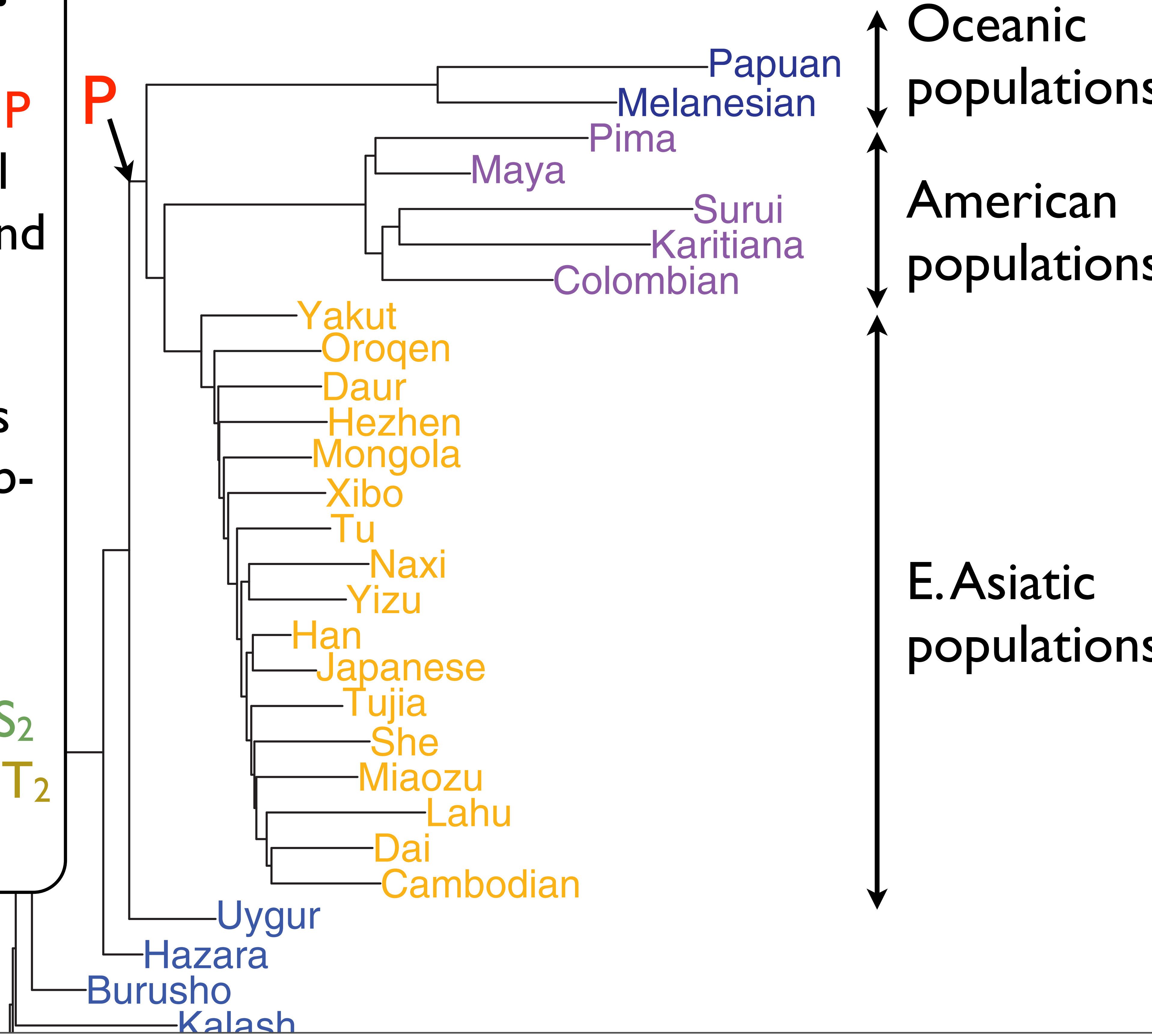


Interpretation:

1. there was a group P of people ancestral to both Oc., Am. and E.A. populations.

2. a separation of this population into sub-populations S_1, S_2

3. second: a further subdivision of the S_2 population into T_1, T_2

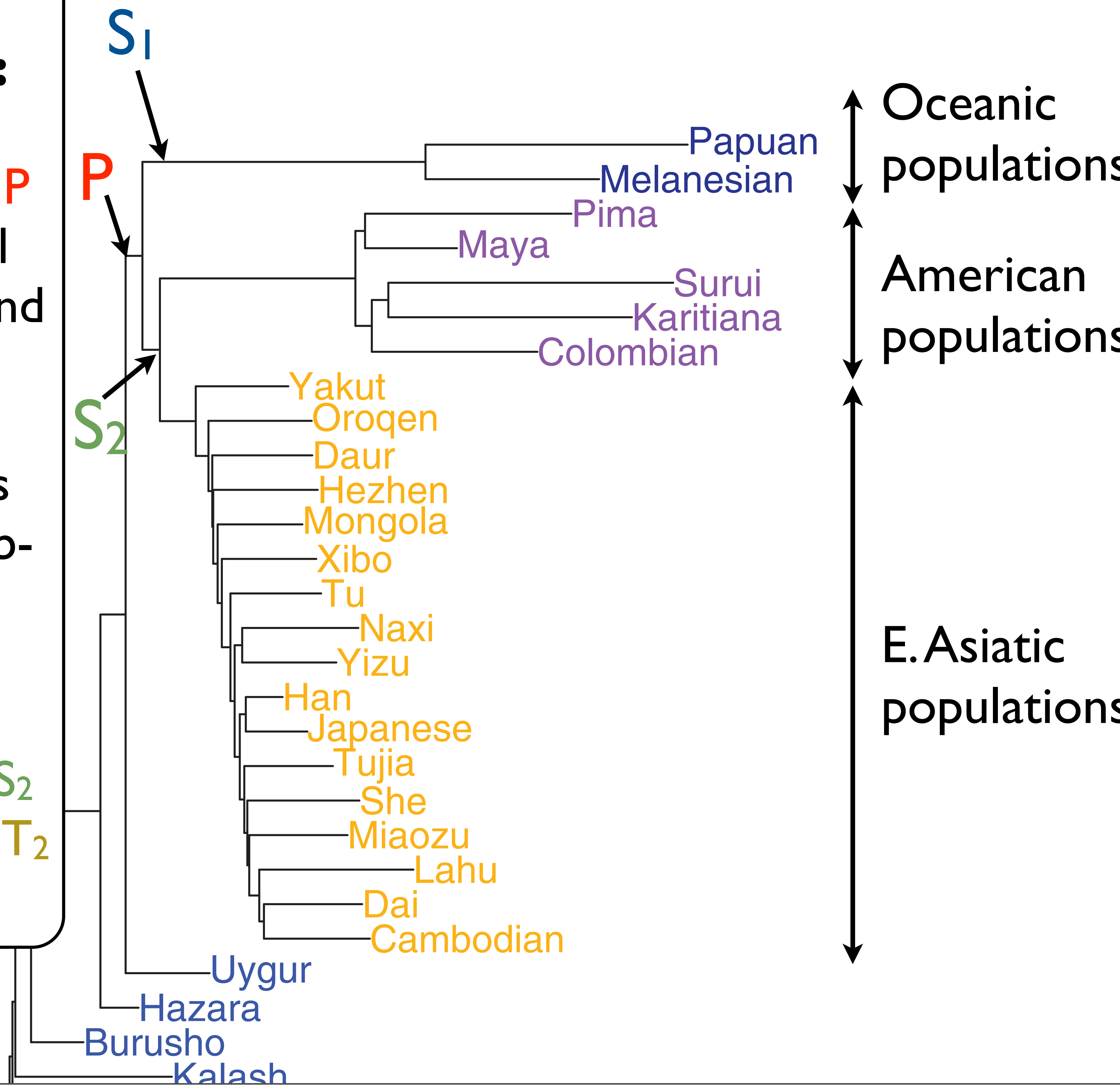


Interpretation:

1. there was a group P of people ancestral to both Oc., Am. and E.A. populations.

2. a separation of this population into sub-populations S_1, S_2

3. second: a further subdivision of the S_2 population into T_1, T_2

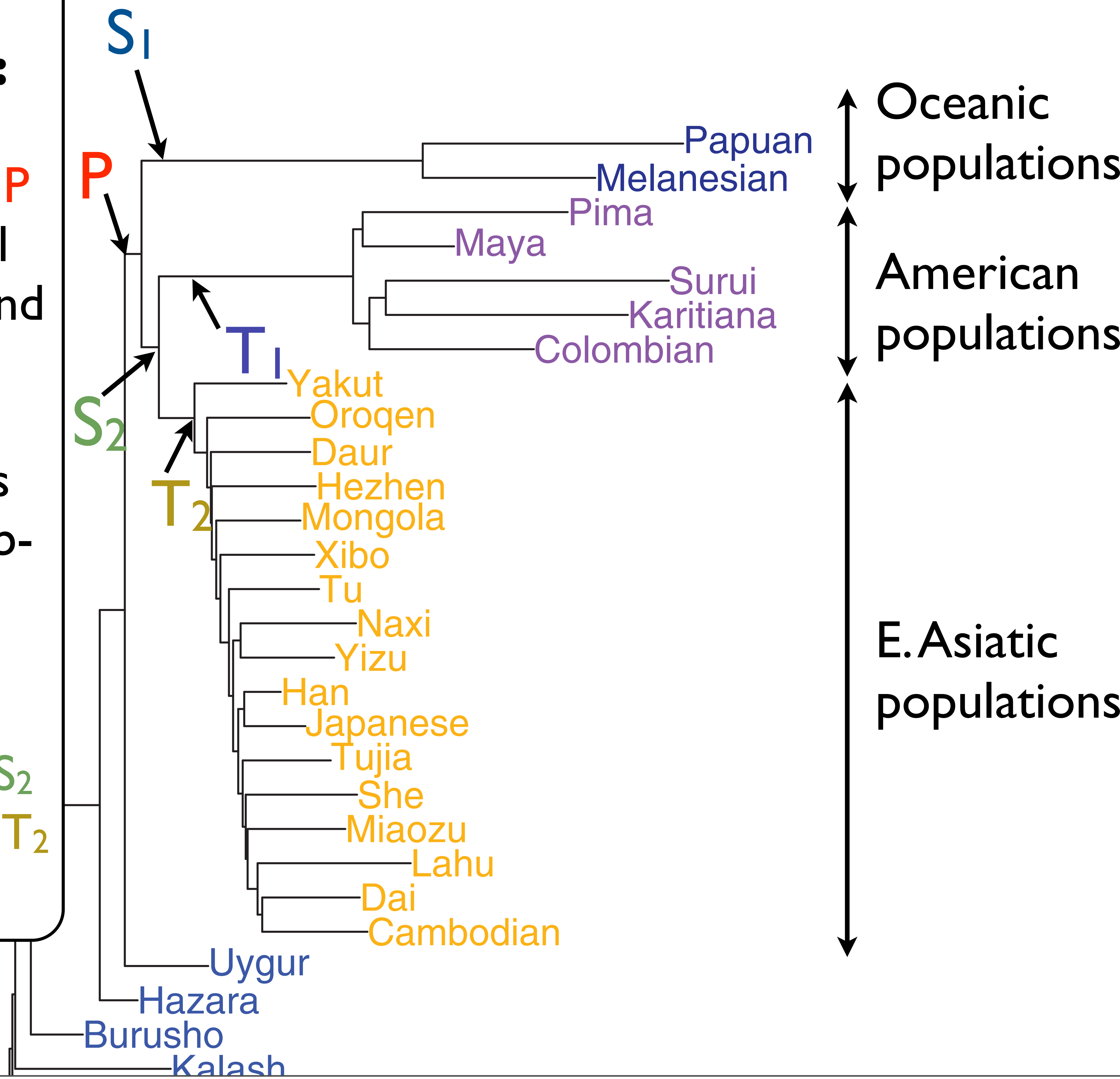


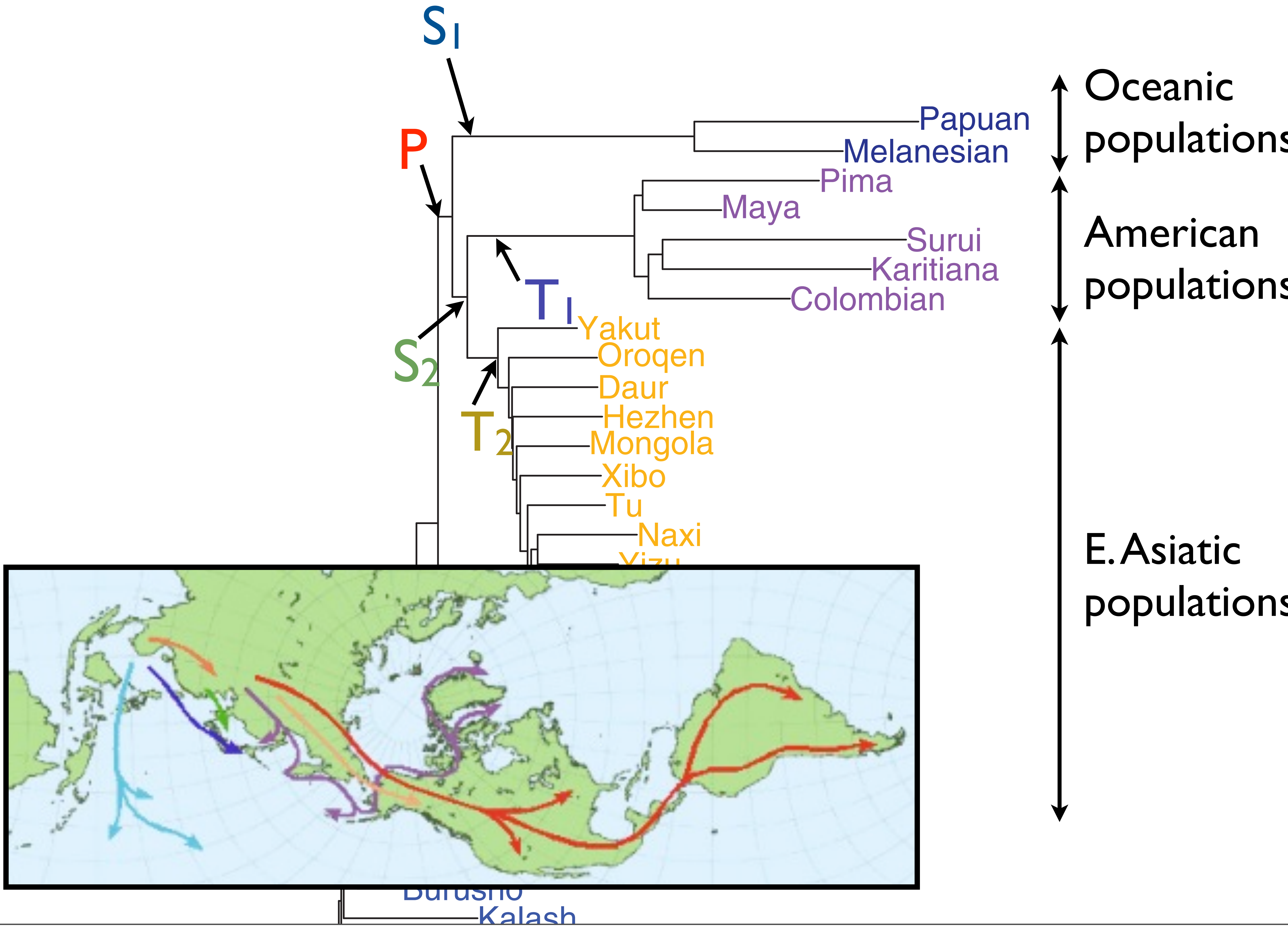
Interpretation:

1. there was a group P of people ancestral to both Oc., Am. and E.A. populations.

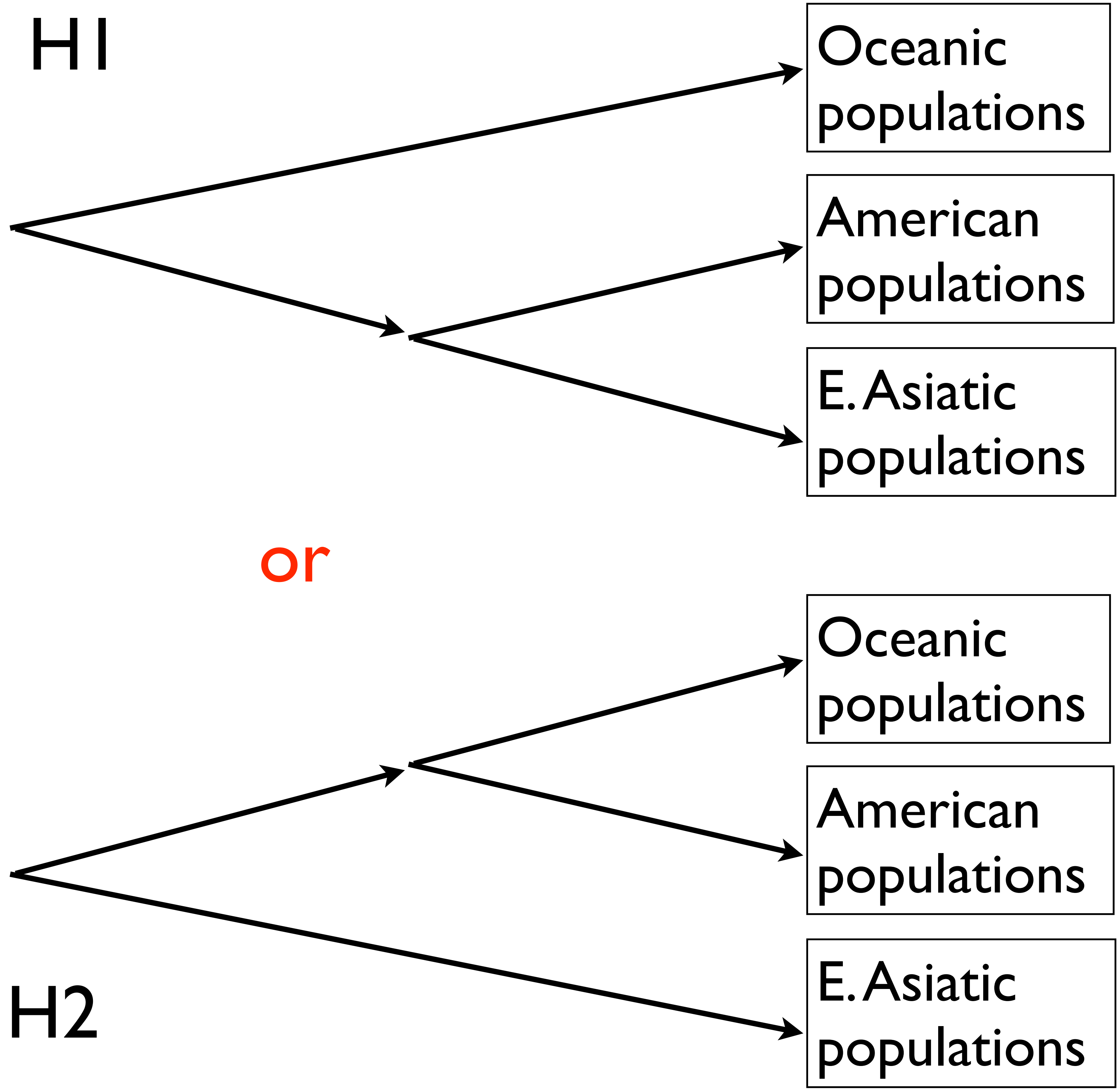
2. a separation of this population into sub-populations S_1, S_2

3. second: a further subdivision of the S_2 population into T_1, T_2




S1
P
T1
S2
T2
Papuan
Melanesian
Pima
Maya
Surui
Karitiana
Colombian
Yakut
Oroqen
Daur
Hezhen
Mongola
Xibo
Tu
Naxi
Kalash
Oceanic populations
American populations
E. Asiatic populations

Simplified example



Can we use the likelihood ratio?

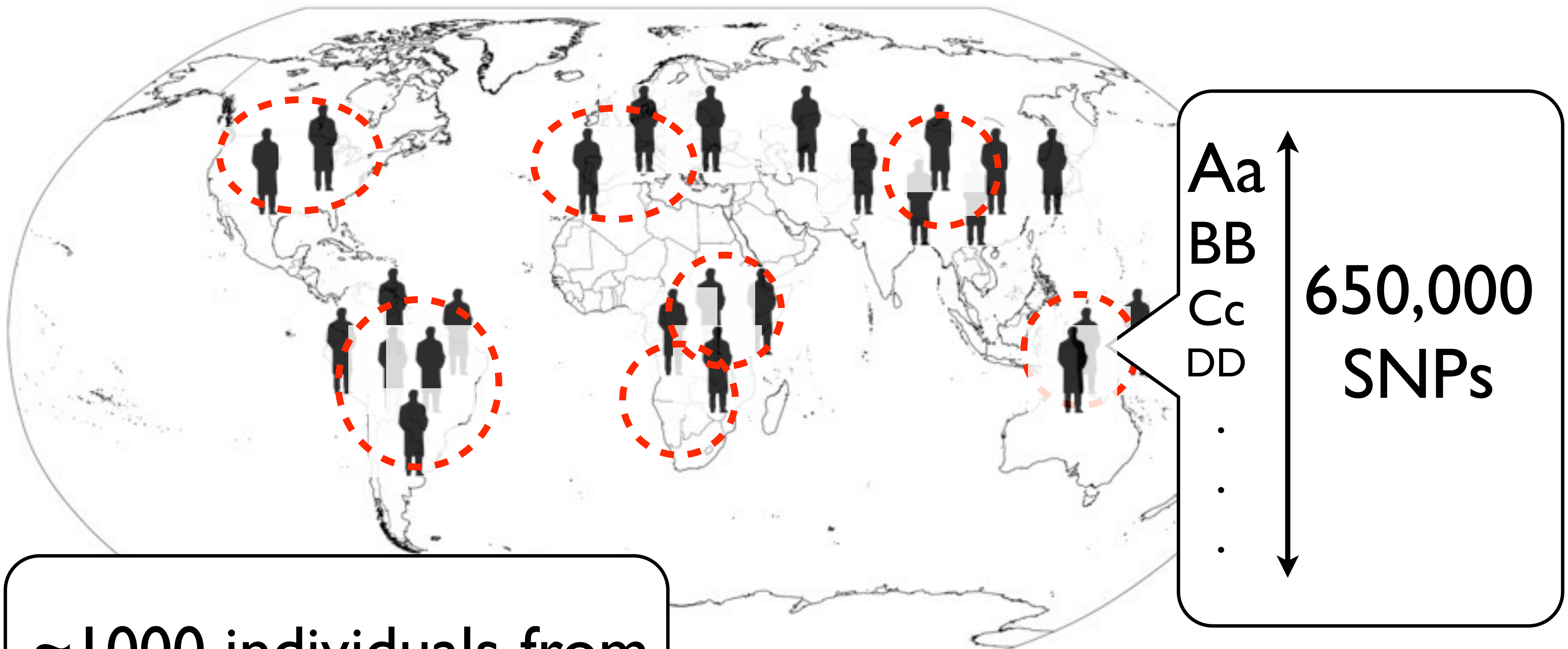
$$\frac{P(Data \mid H1)}{P(Data \mid H2)}$$

Questions:

What is the data?

What is the model, i.e. what is P

Data: first type



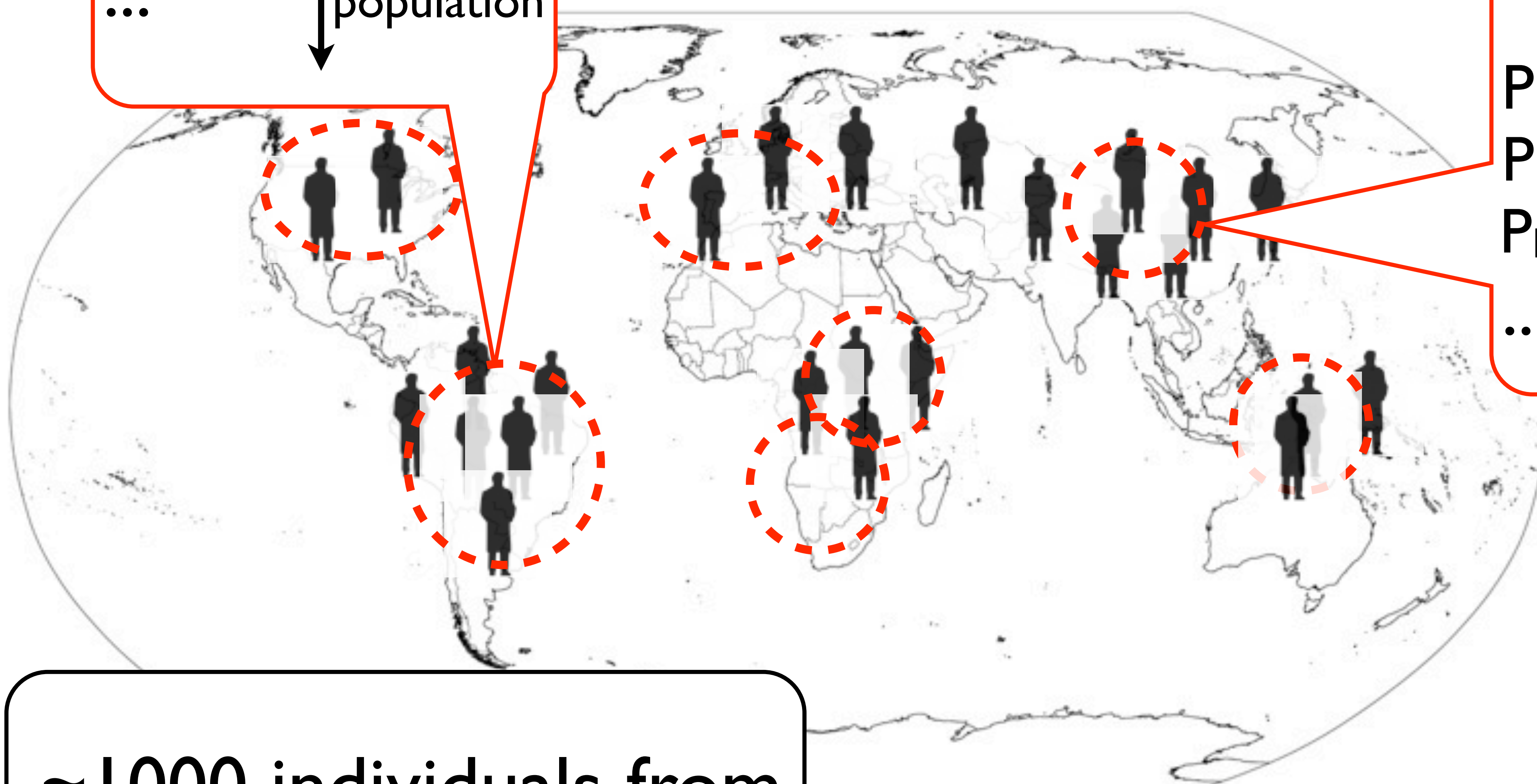
~1000 individuals from
~50 populations

Human Genome Diversity Panel

Data

$P_{Maya}(A)$
$P_{Maya}(B)$
$P_{Maya}(C)$
...

Compute allele frequency for each population



$P_{Han}(A)$
$P_{Han}(B)$
$P_{Han}(C)$
...

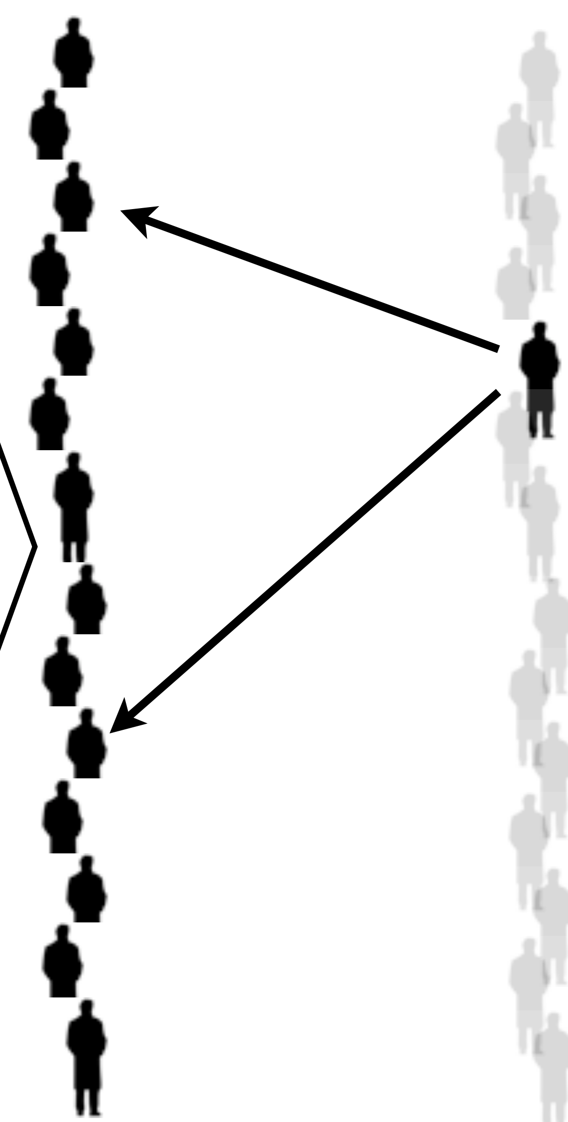
Different because of finite pop., non-uniform mixing

~1000 individuals from ~50 populations

Human Genome Diversity Panel

Model 1: Wright-Fisher

Generation 1 Generation 2



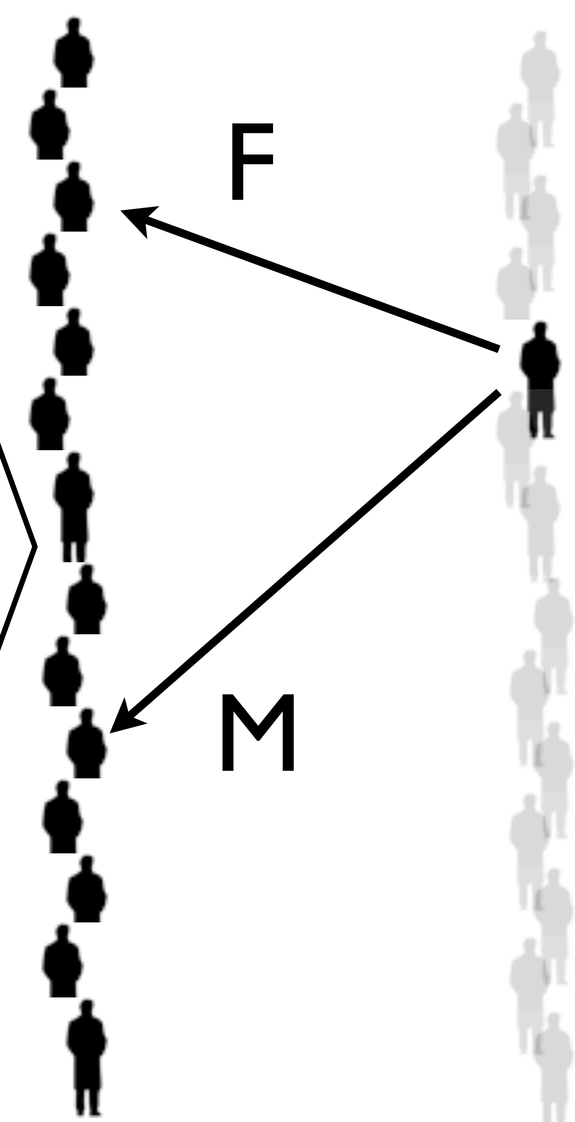
Suppose there are 50 people in the first generation

Random mating: Assume each individual in the next generation has a father taken uniformly at random from the previous generation, and a mother taken independently at random

Model 1: Wright-Fisher

Generation 1 | Generation 2

Suppose there are:
70 copies of the A allele,
30 copies of the a allele

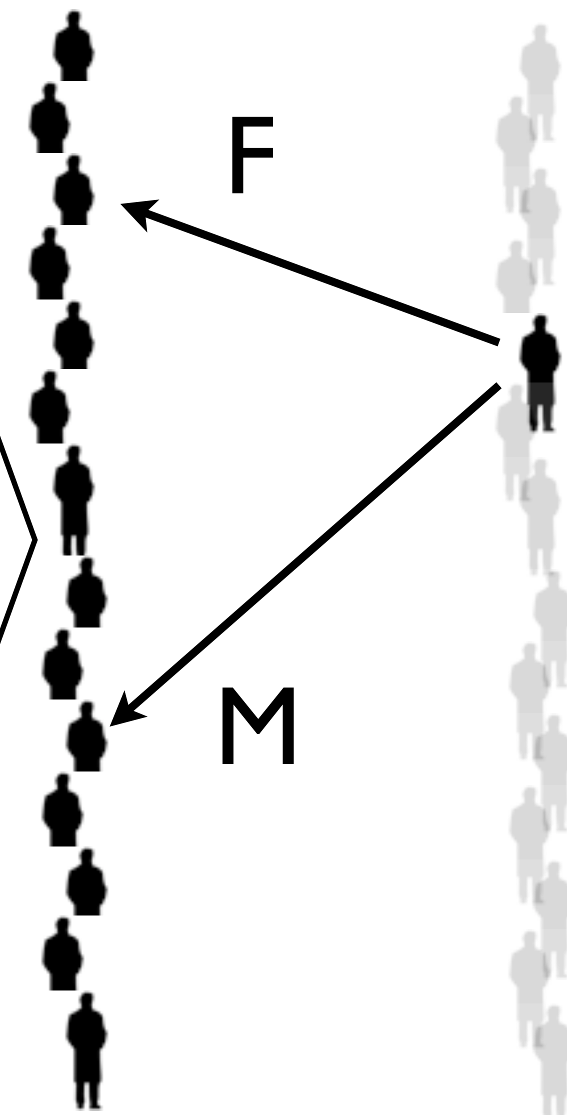


What is the probability that the allele inherited from the father is A?

Model 1: Wright-Fisher

Generation 1 Generati

Suppose there are still 50 peoples (100 allele copies) at generation 2. What is your best guess for the number of copies of the A allele in generation 2?



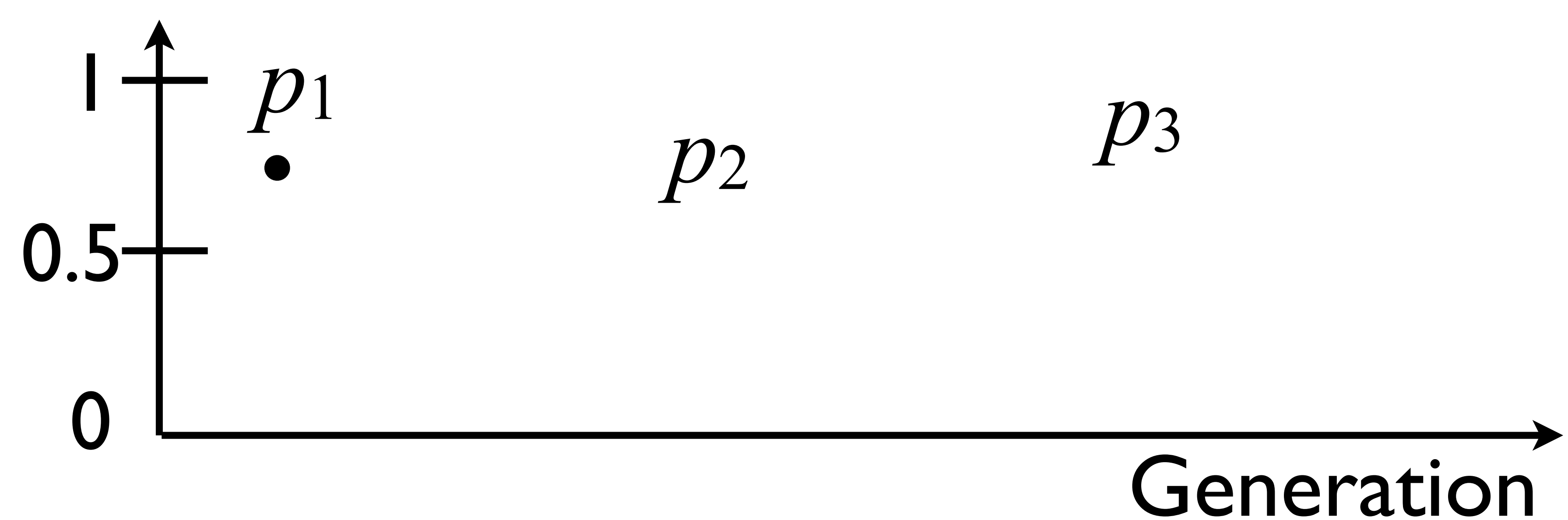
Suppose there are:
70 copies of the A allele,
30 copies of the a allele

Martingale

Generation 1

Suppose there are initially:
70 copies of the A allele,
30 copies of the a allele

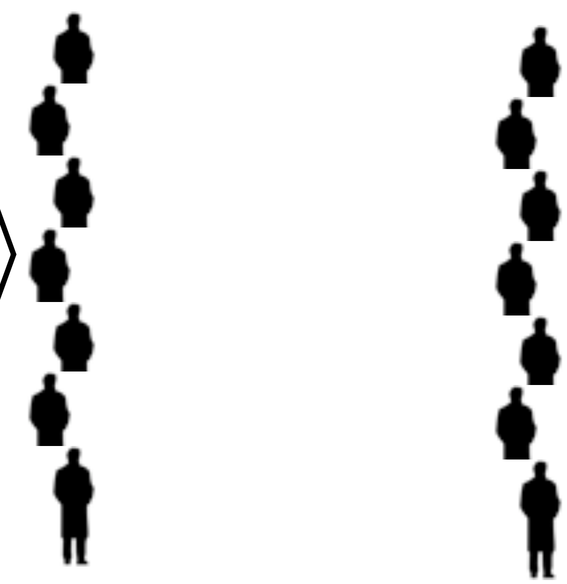
Fraction of the pop. with the A allele = P(A)



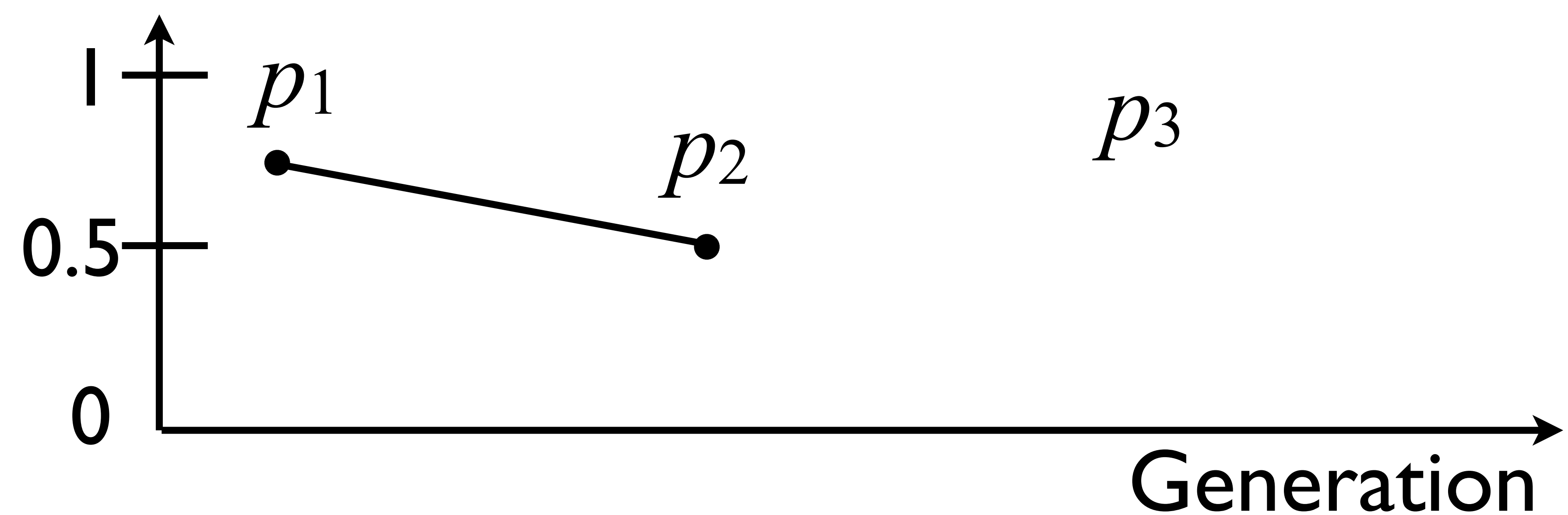
Martingale

Generation 1   Generation 2

Suppose there are initially:
70 copies of the A allele,
30 copies of the a allele



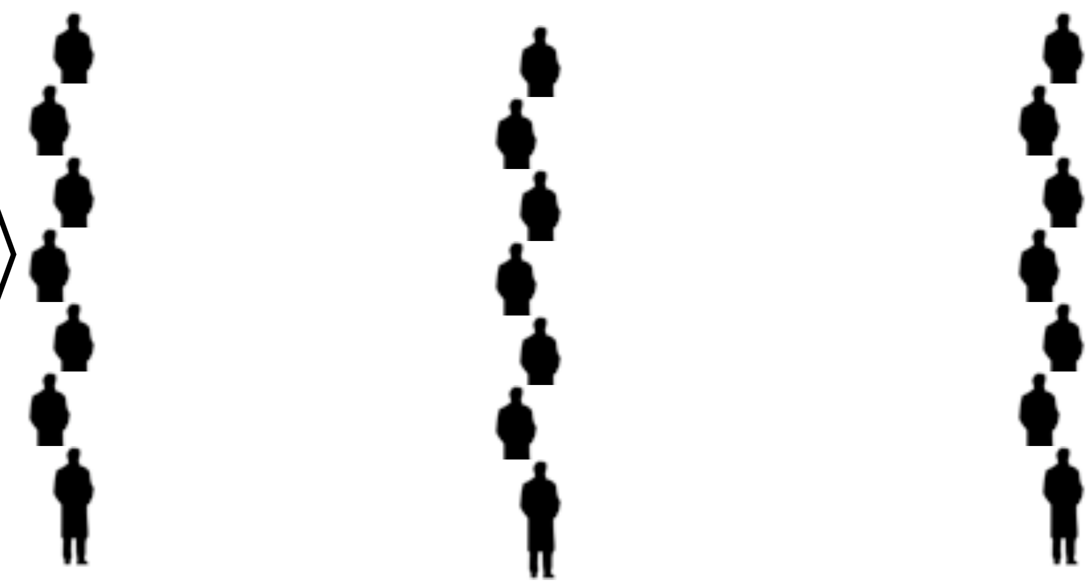
Fraction of the pop. with the A allele = P(A)



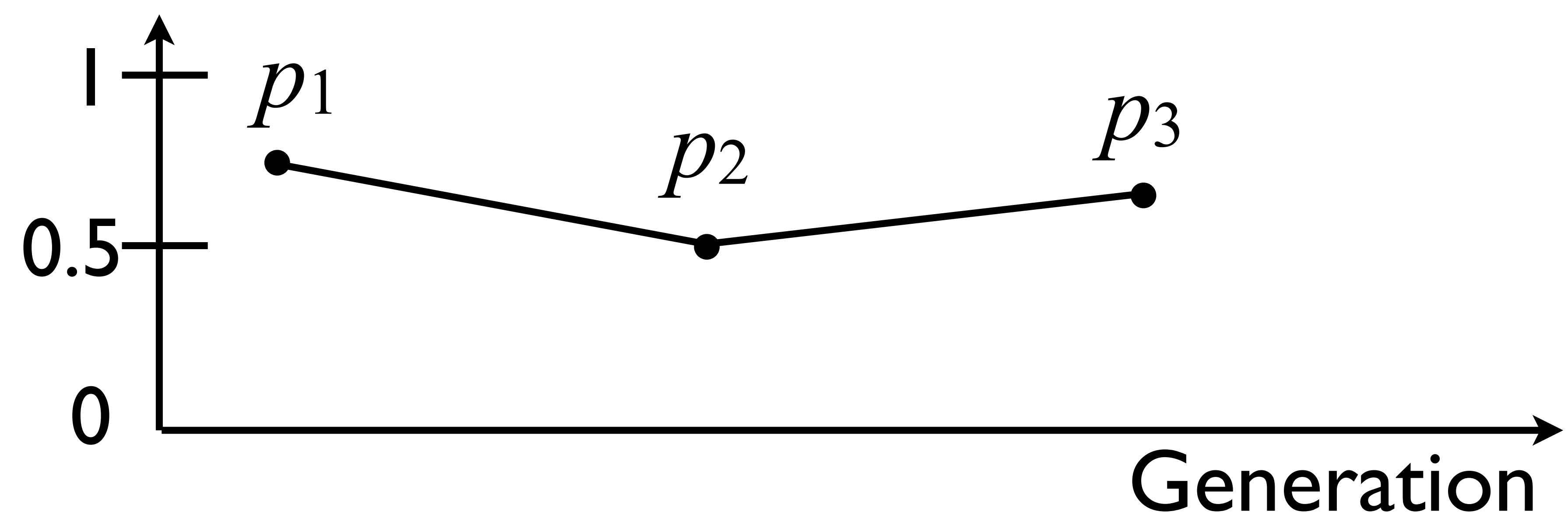
Martingale

Generation 1 Generation 2 Generation 3

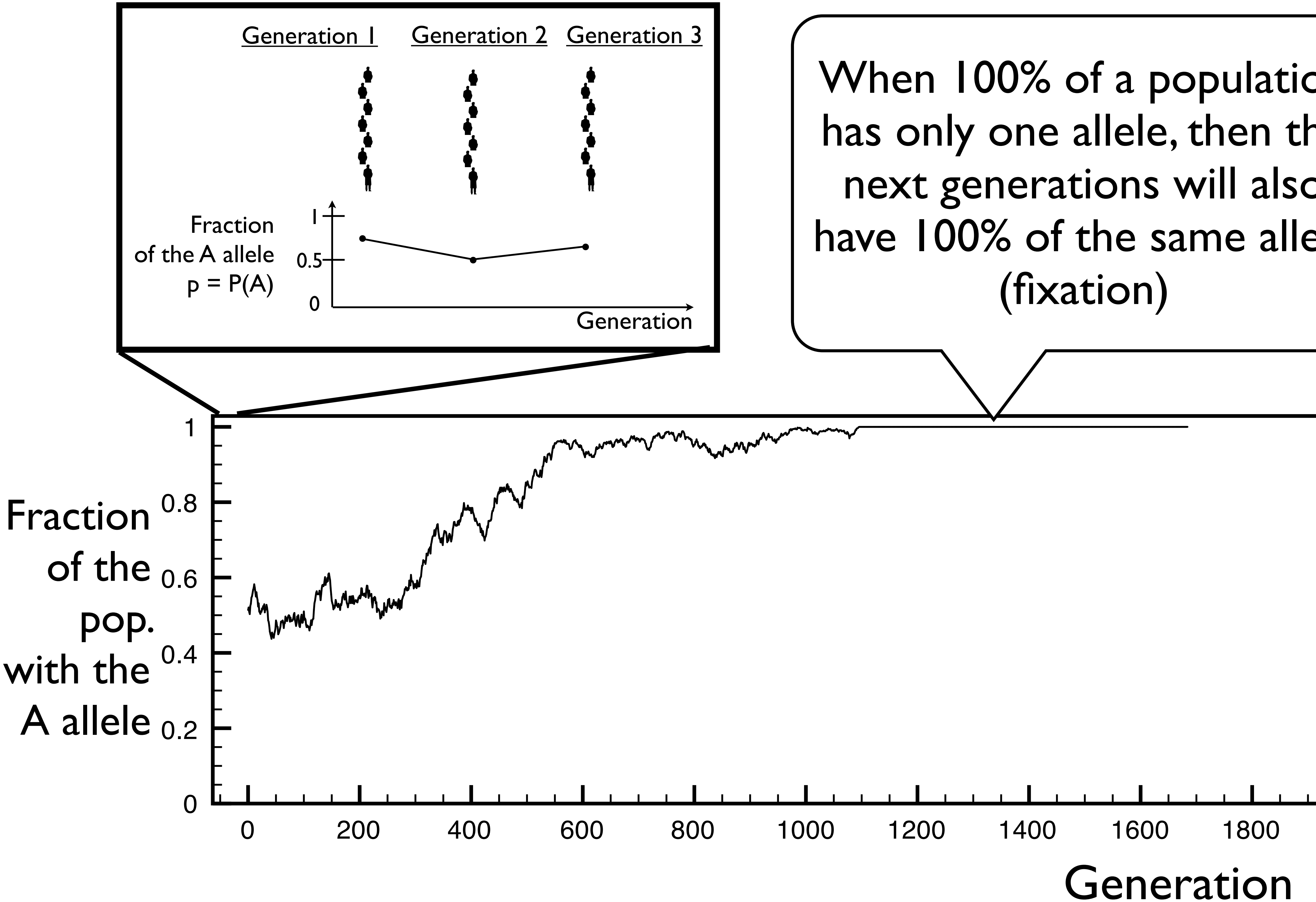
Suppose there are initially:
70 copies of the A allele,
30 copies of the a allele



Fraction of the pop. with the A allele = P(A)



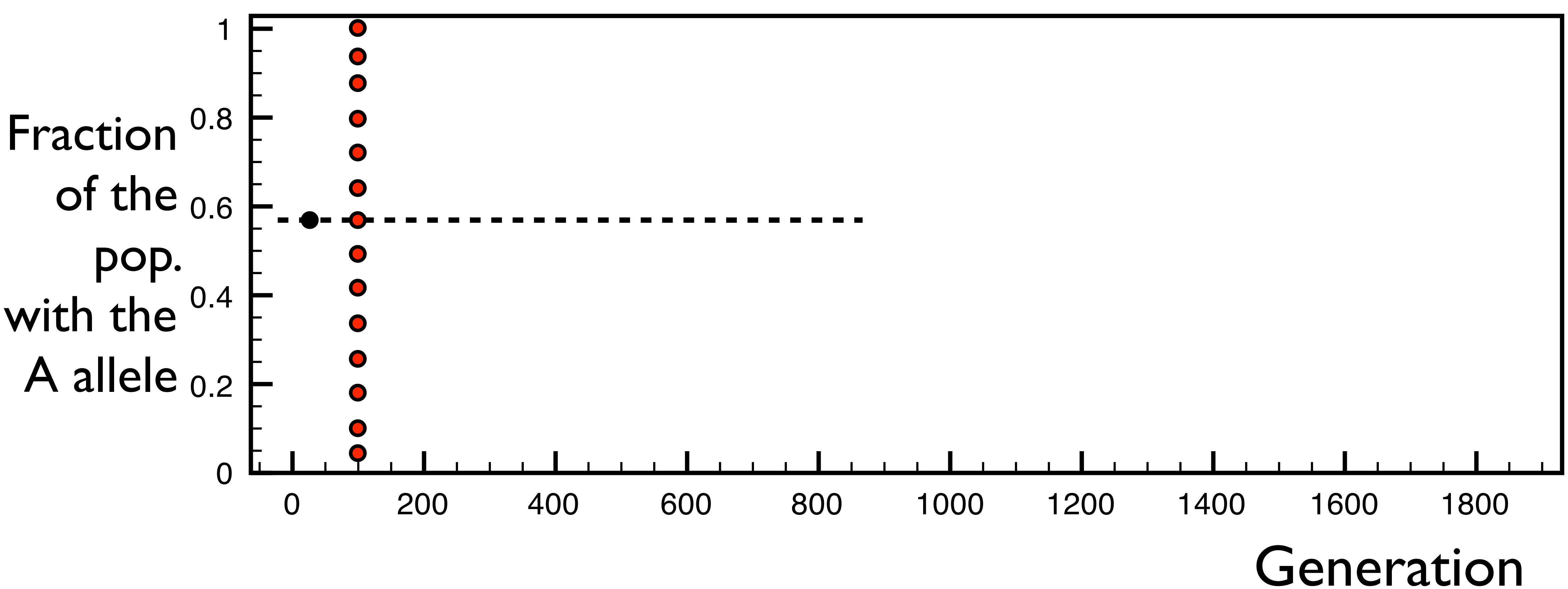
Wright-Fisher model



When 100% of a population has only one allele, then the next generations will also have 100% of the same allele (fixation)

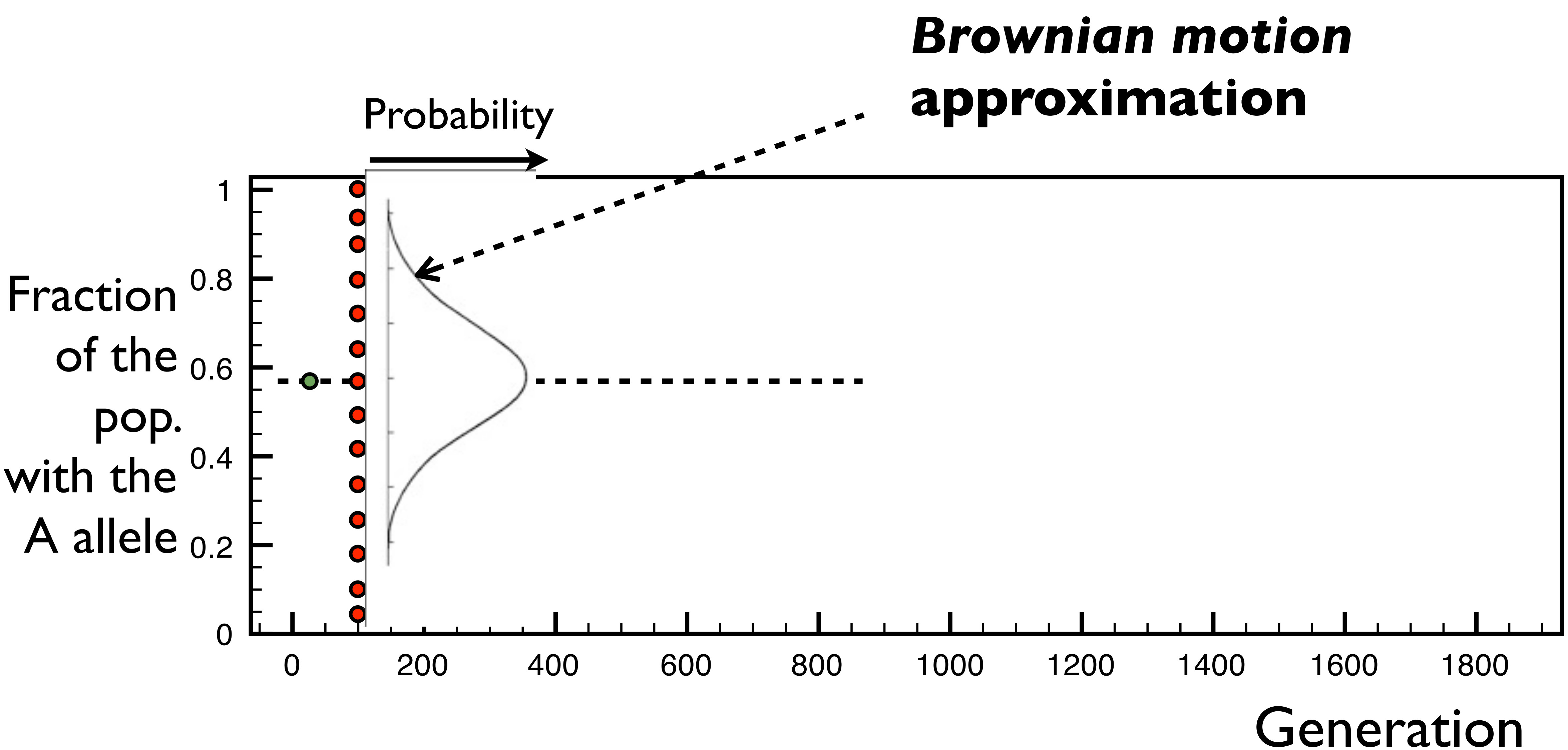
Wright-Fisher model

If the allele frequency is 0.5 initially, what is the probability distribution over allele frequencies after 100 generations?



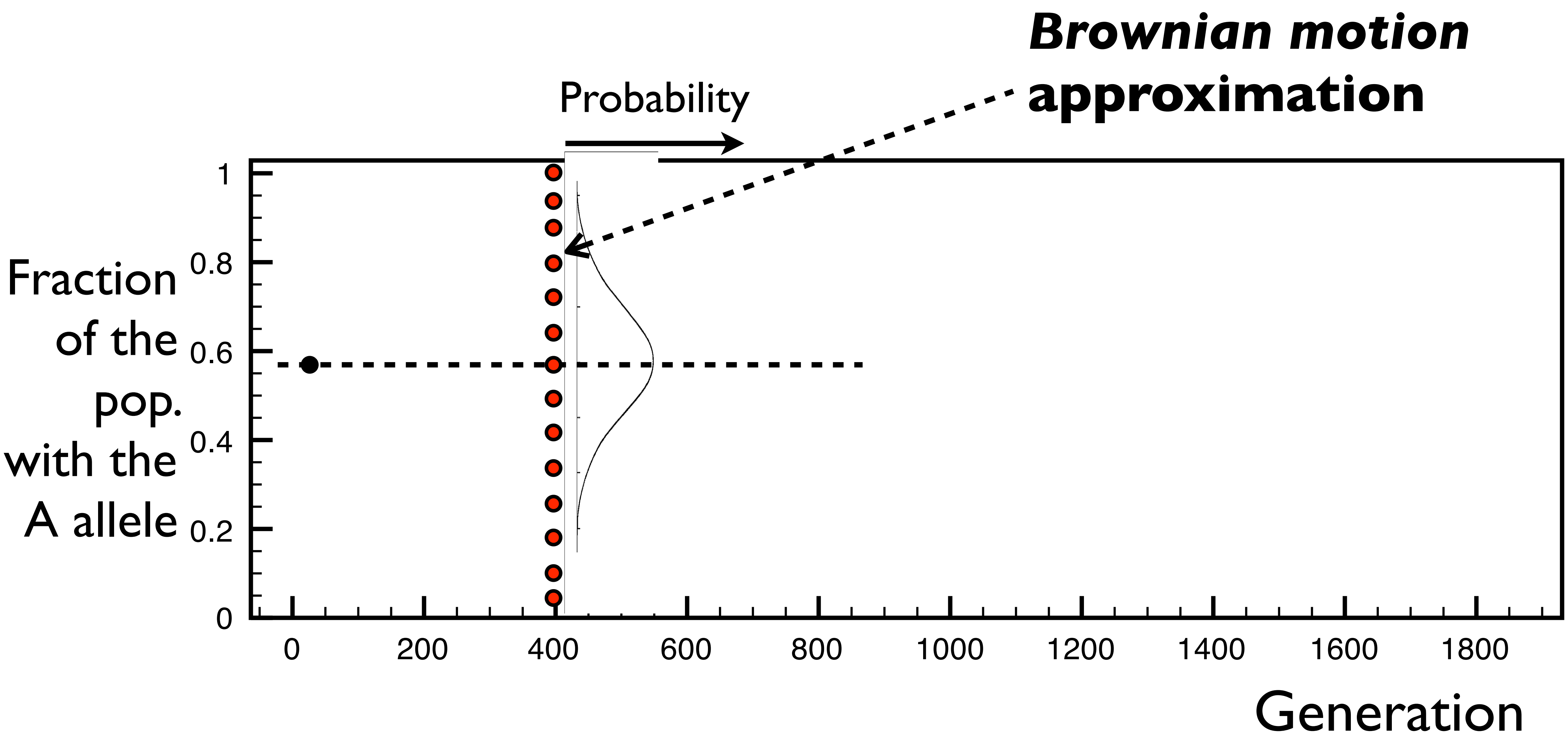
Wright-Fisher model

If the allele frequency is 0.5 initially, what is the probability distribution over allele frequencies after 100 generations?

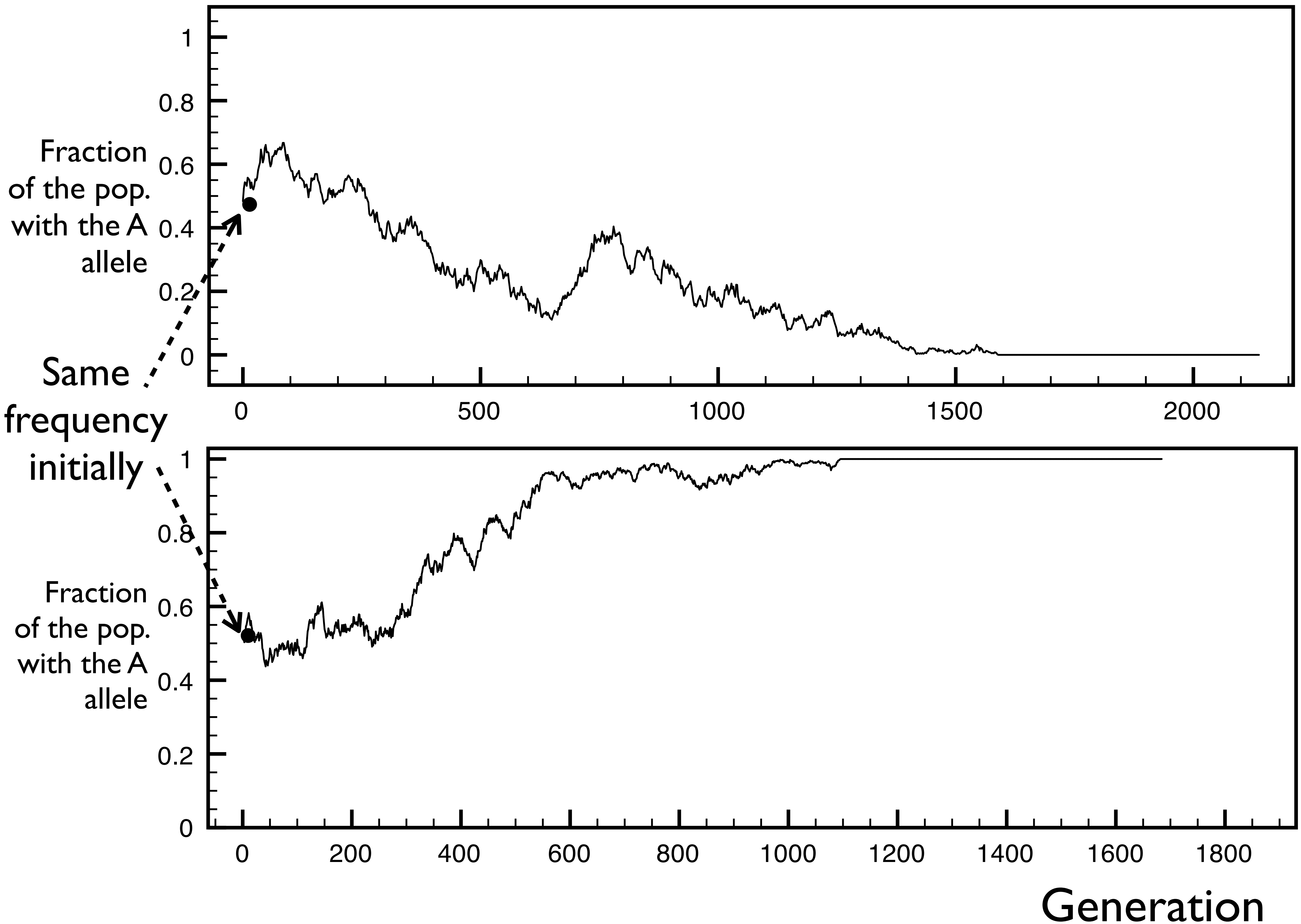


Wright-Fisher model

If the allele frequency is 0.5 initially, what is the probability distribution over allele frequencies after 100 generations?

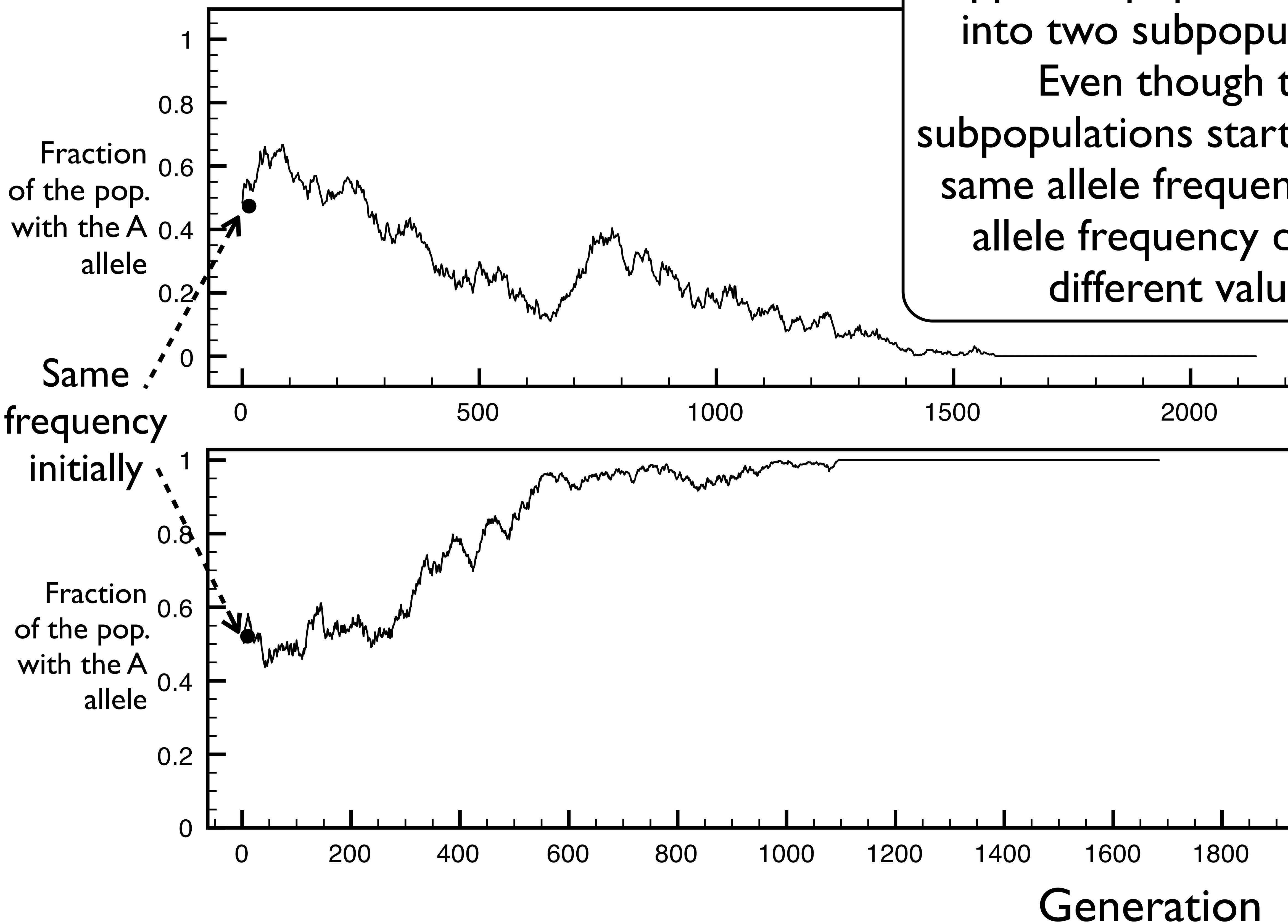


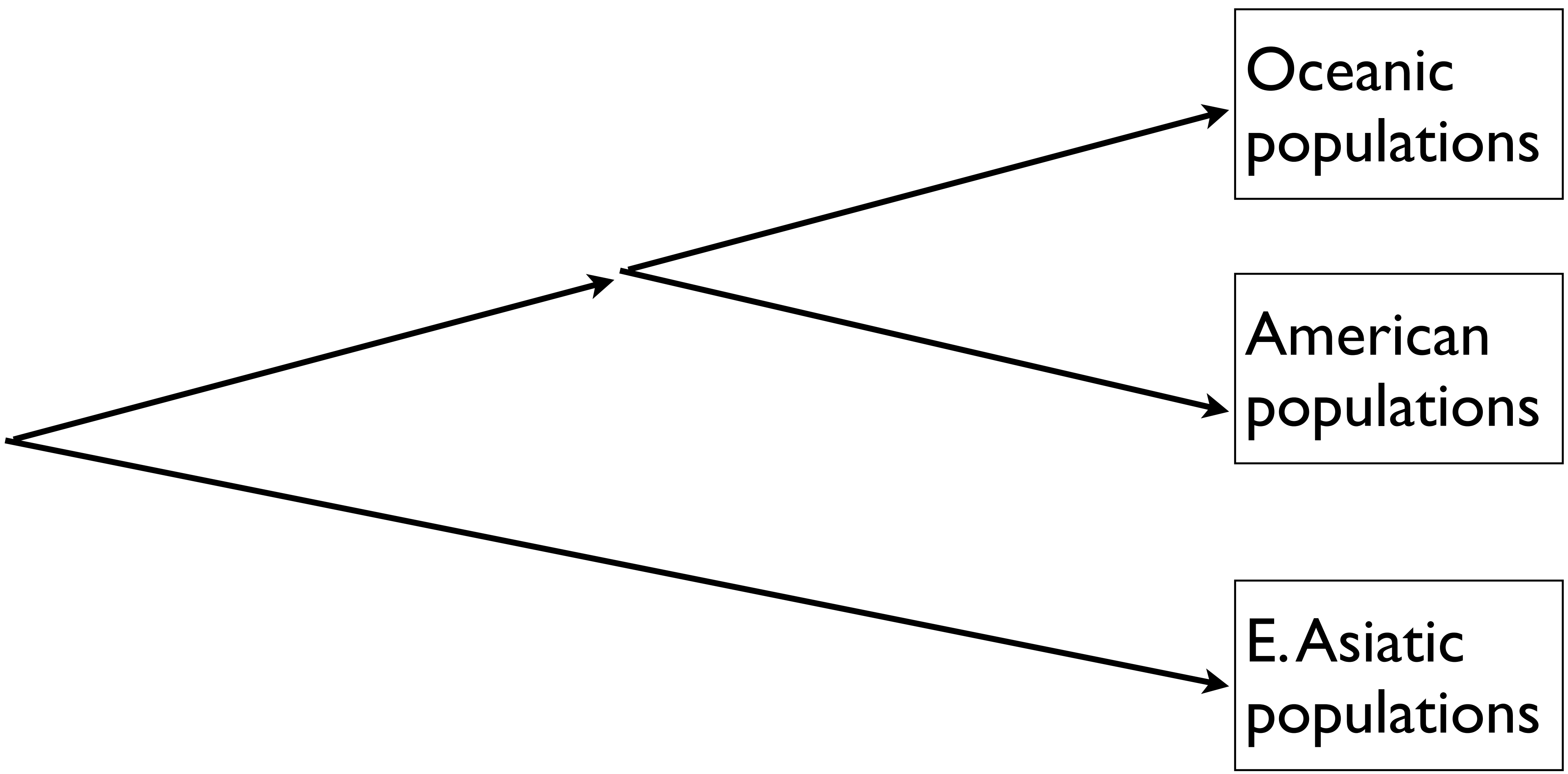
Wright-Fisher model

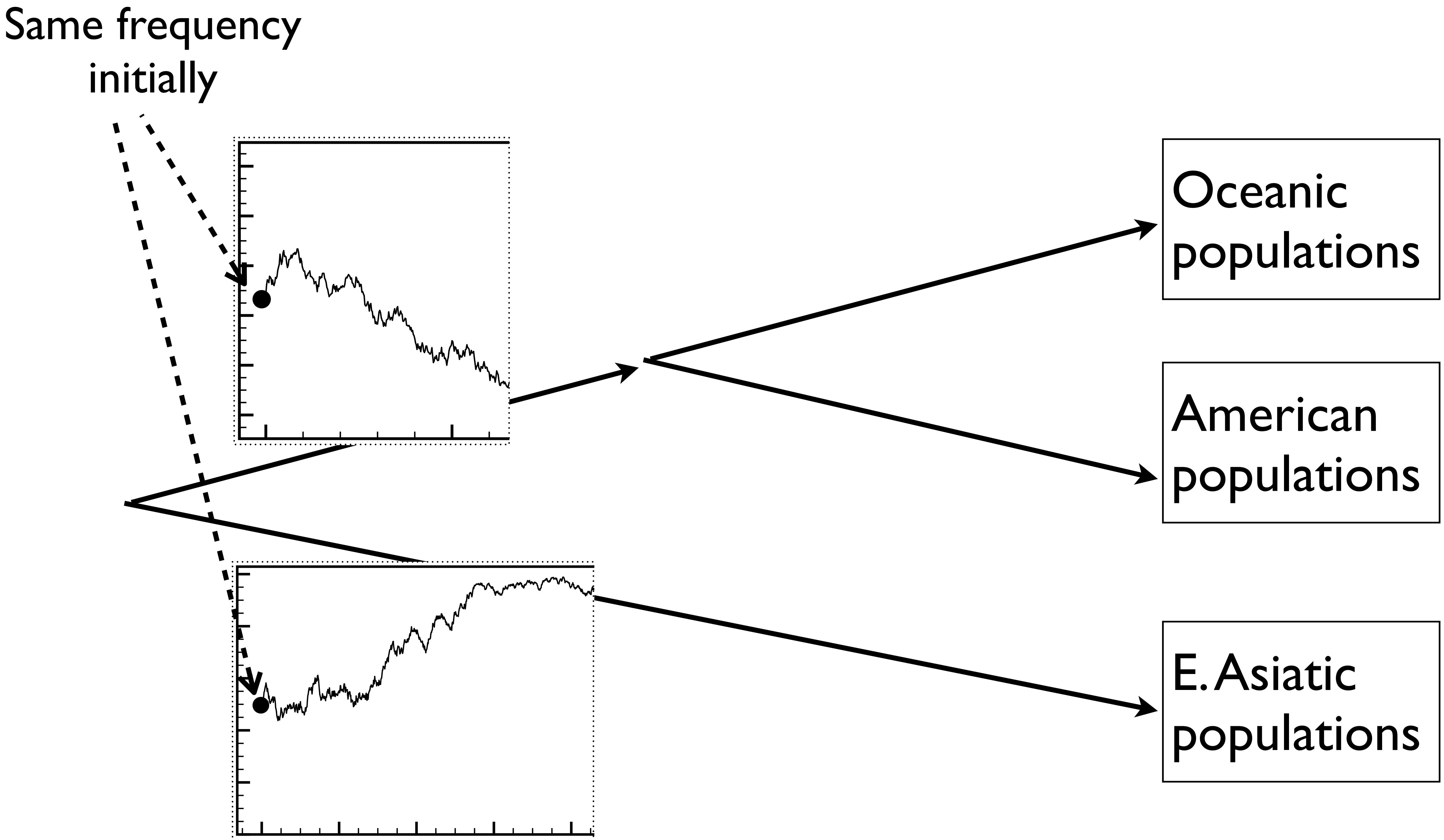


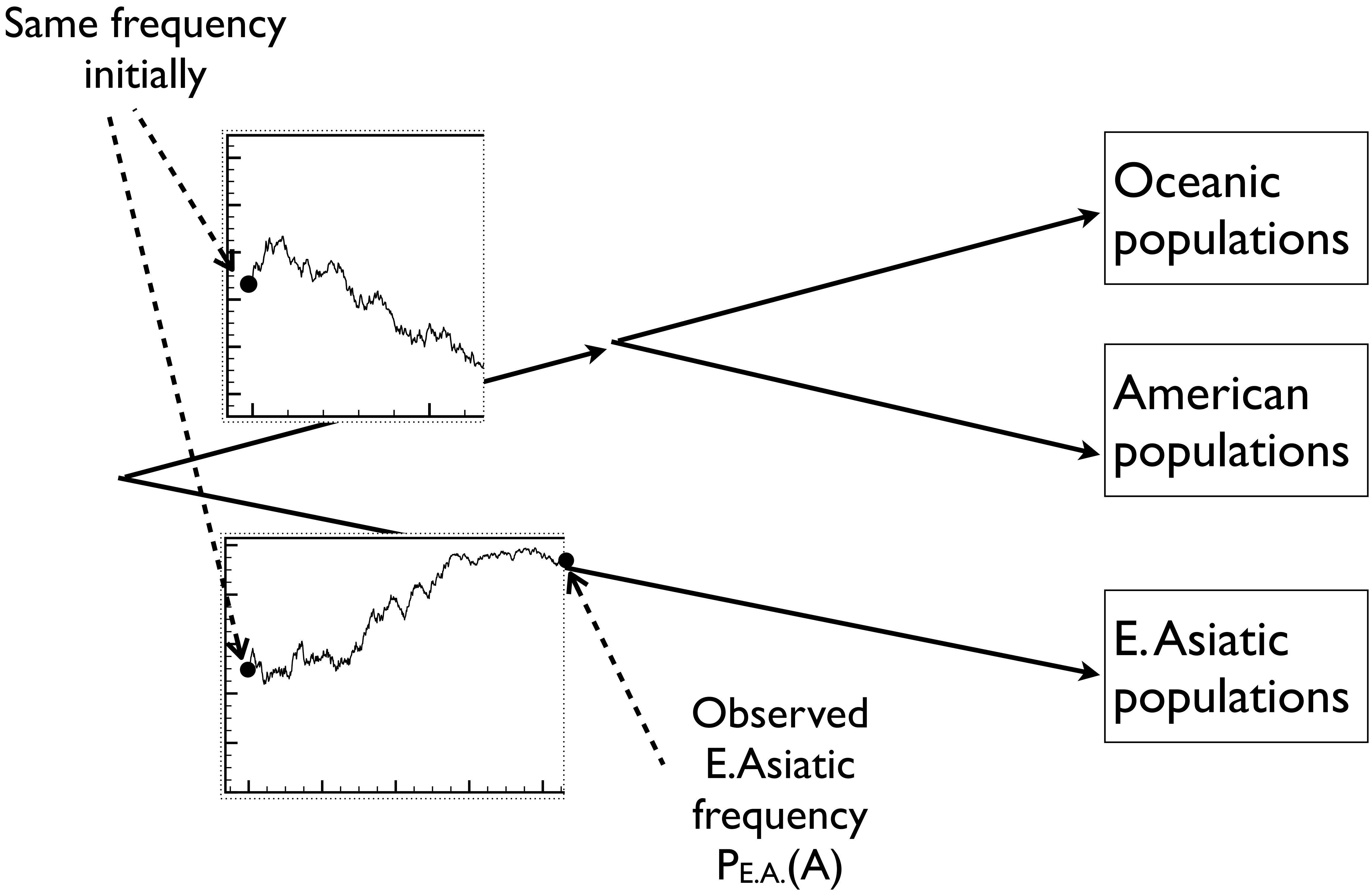
Wright-Fisher model

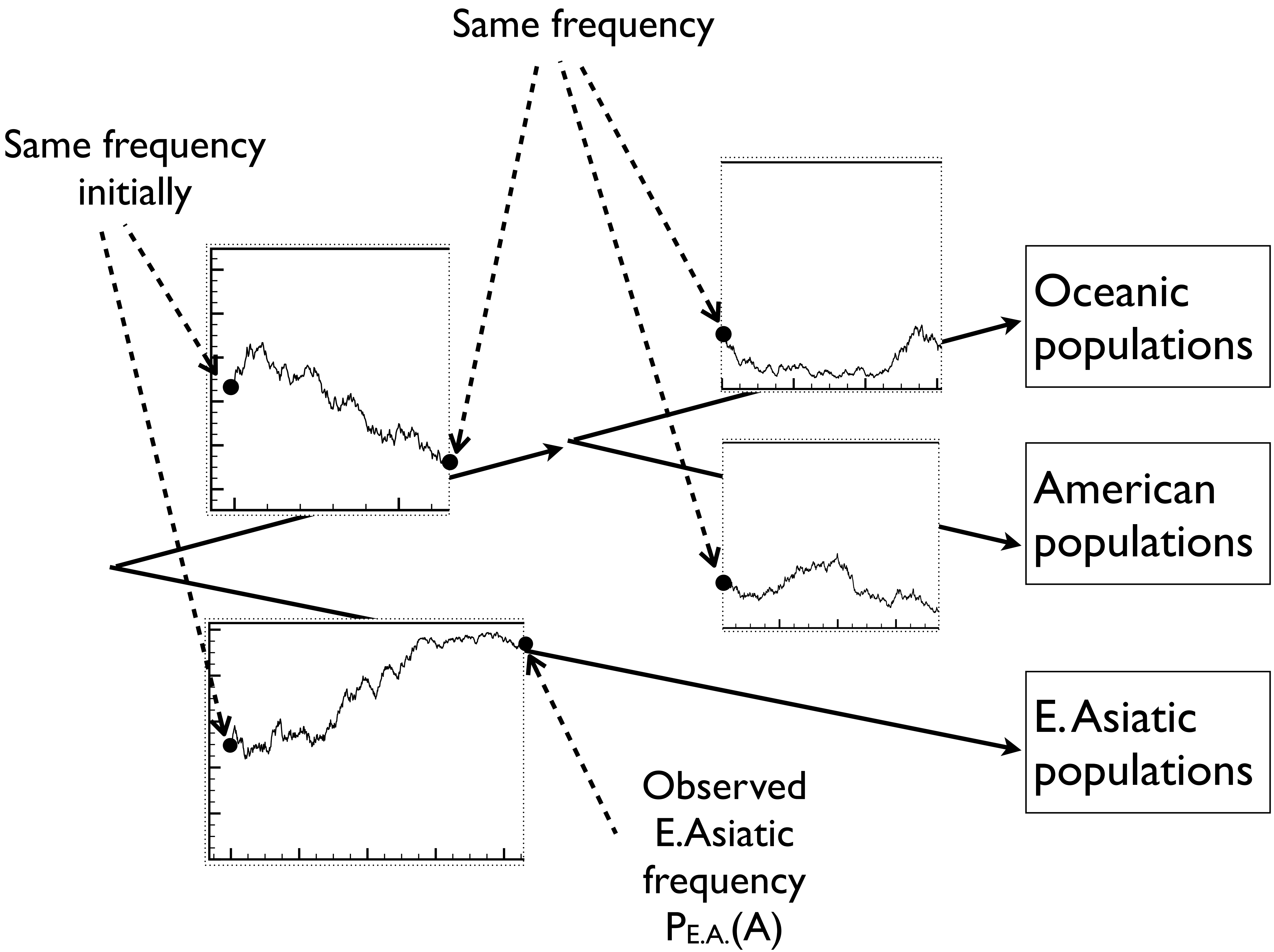
Suppose a population is split into two subpopulations. Even though the subpopulations start with the same allele frequency their allele frequency drift to different values

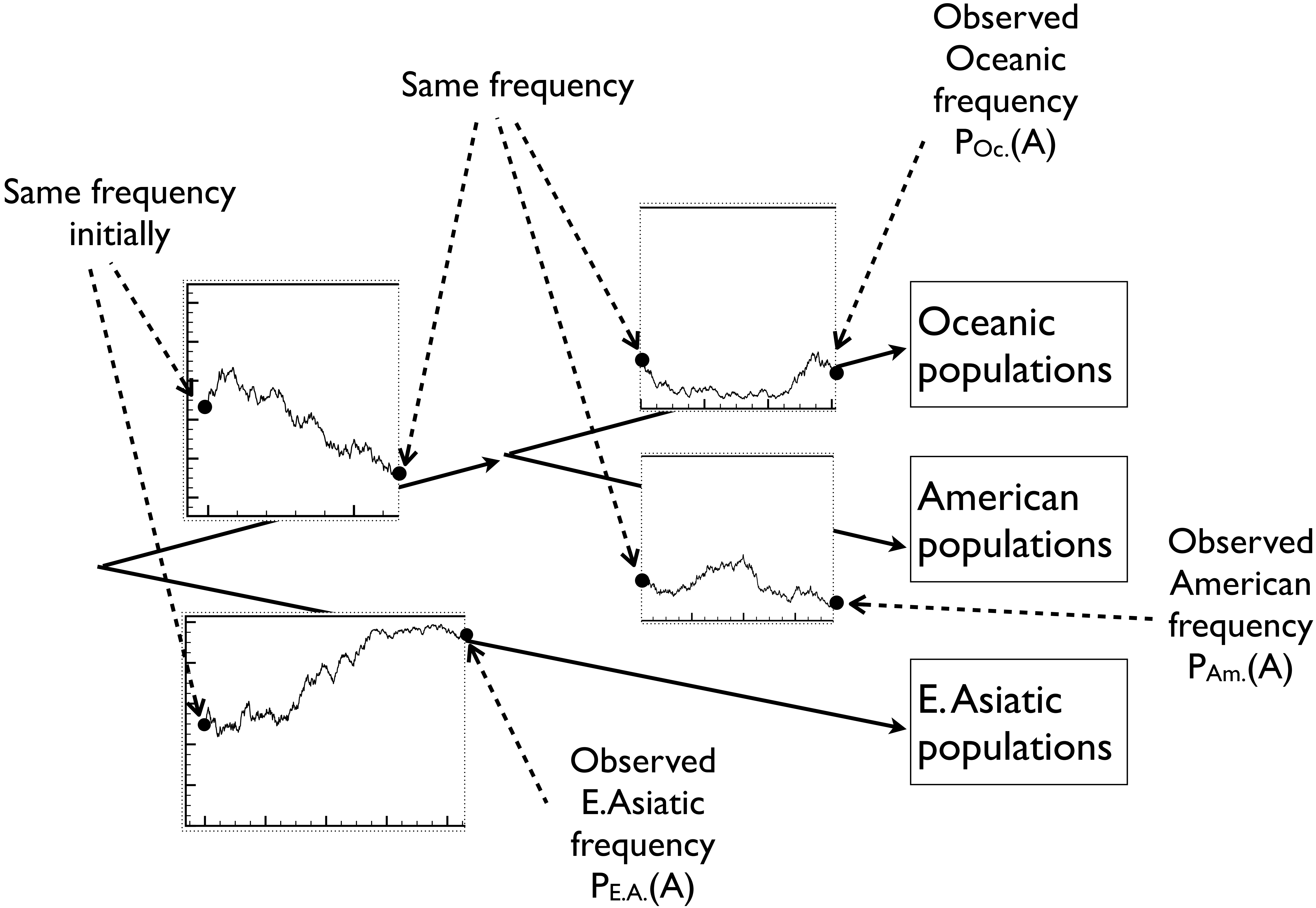


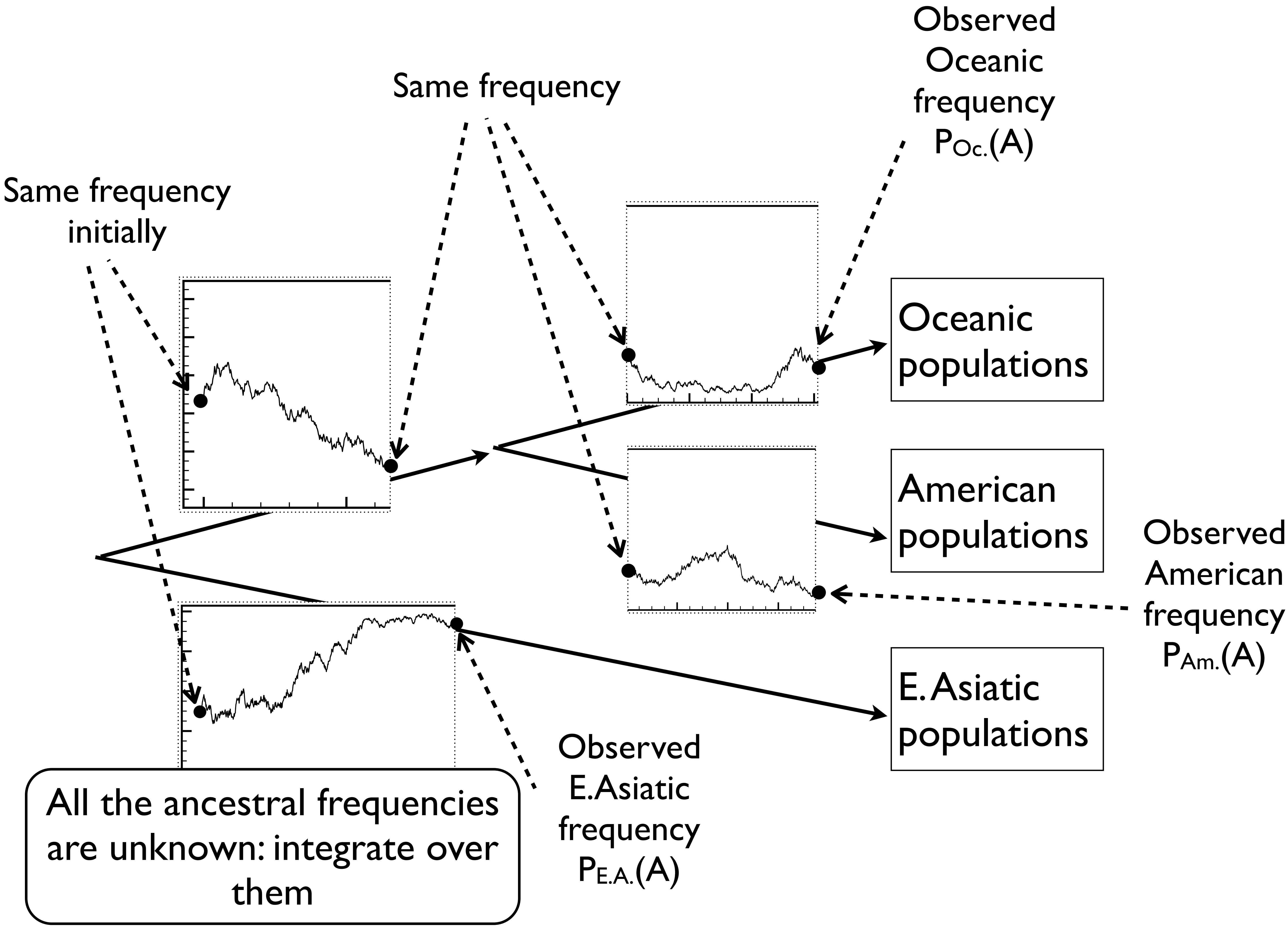

Oceanic populations
American populations
E. Asiatic populations

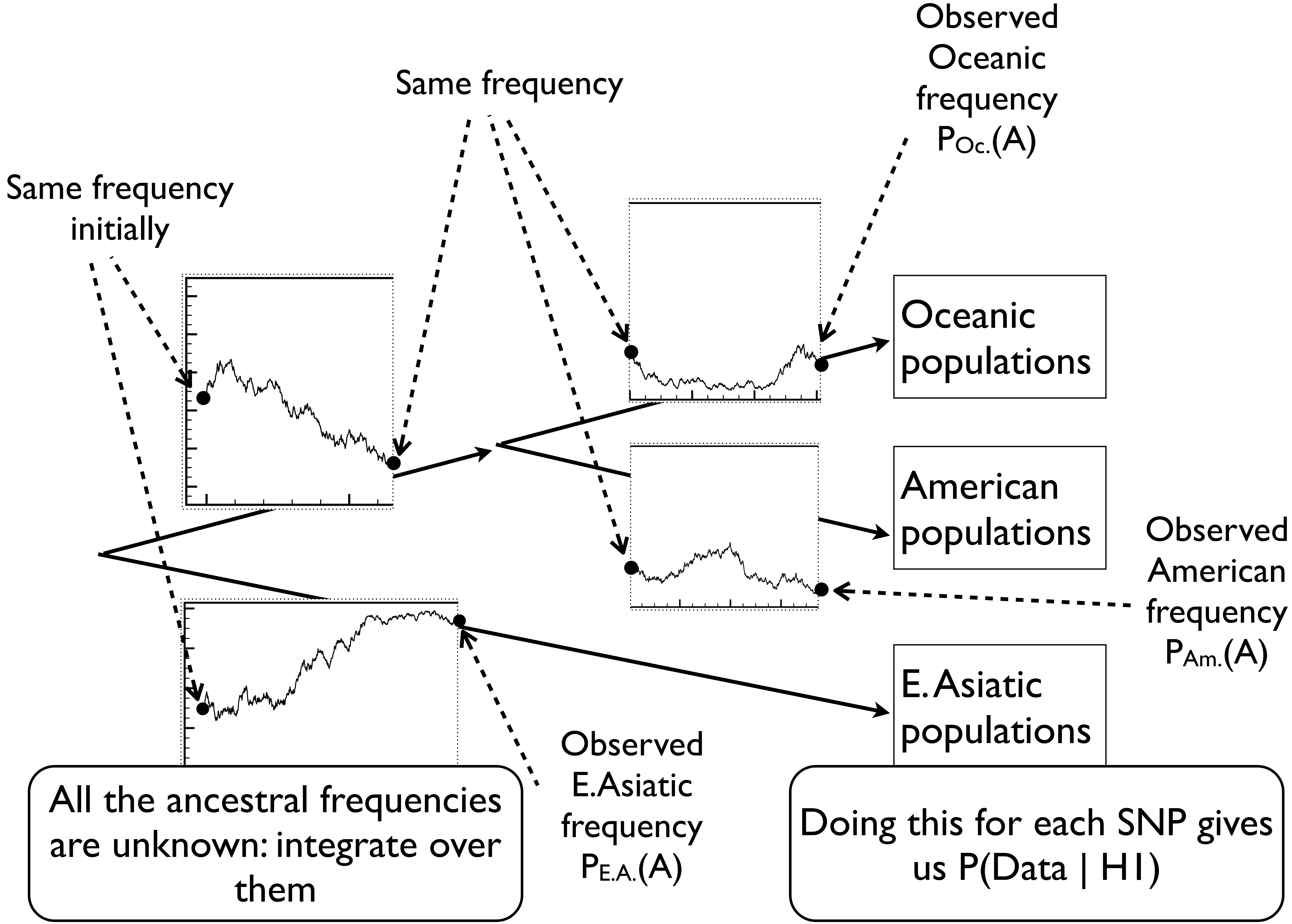

Same frequency
initially
Oceanic
populations
American
populations
E. Asiatic
populations


Same frequency initially
Oceanic populations
American populations
E. Asiatic populations
Observed E.Asiatic frequency $P_{E.A.}(A)$

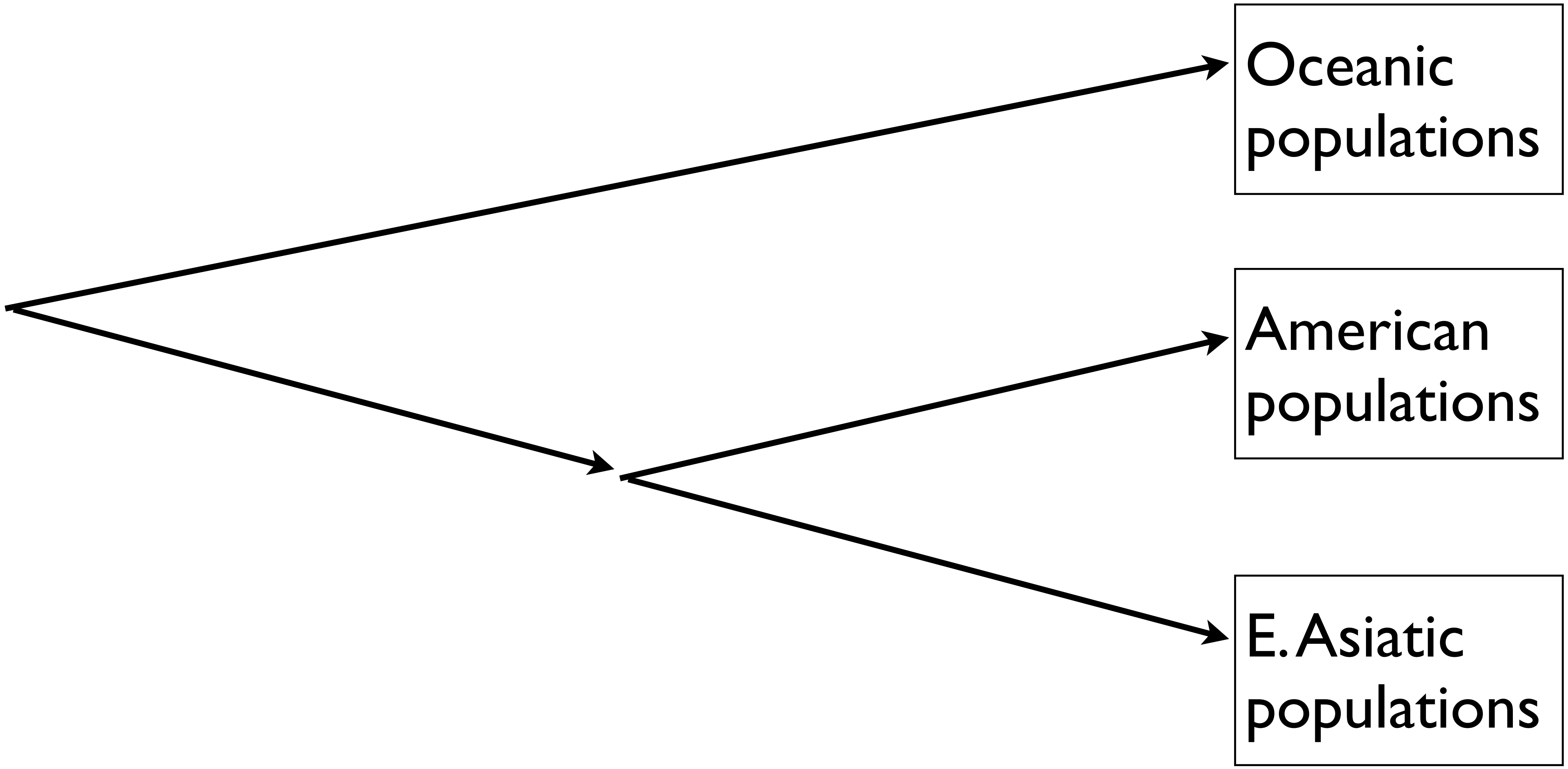

Same frequency
Same frequency
initially
Oceanic
populations
American
populations
E. Asiatic
populations
Observed
E.Asiatic
frequency
$P_{E.A.}(A)$


Same frequency
Observed
Oceanic
frequency
$P_{Oc.}(A)$
Same frequency
initially
Oceanic
populations
American
populations
Observed
American
frequency
$P_{Am.}(A)$
E. Asiatic
populations
Observed
E.Asiatic
frequency
$P_{E.A.}(A)$

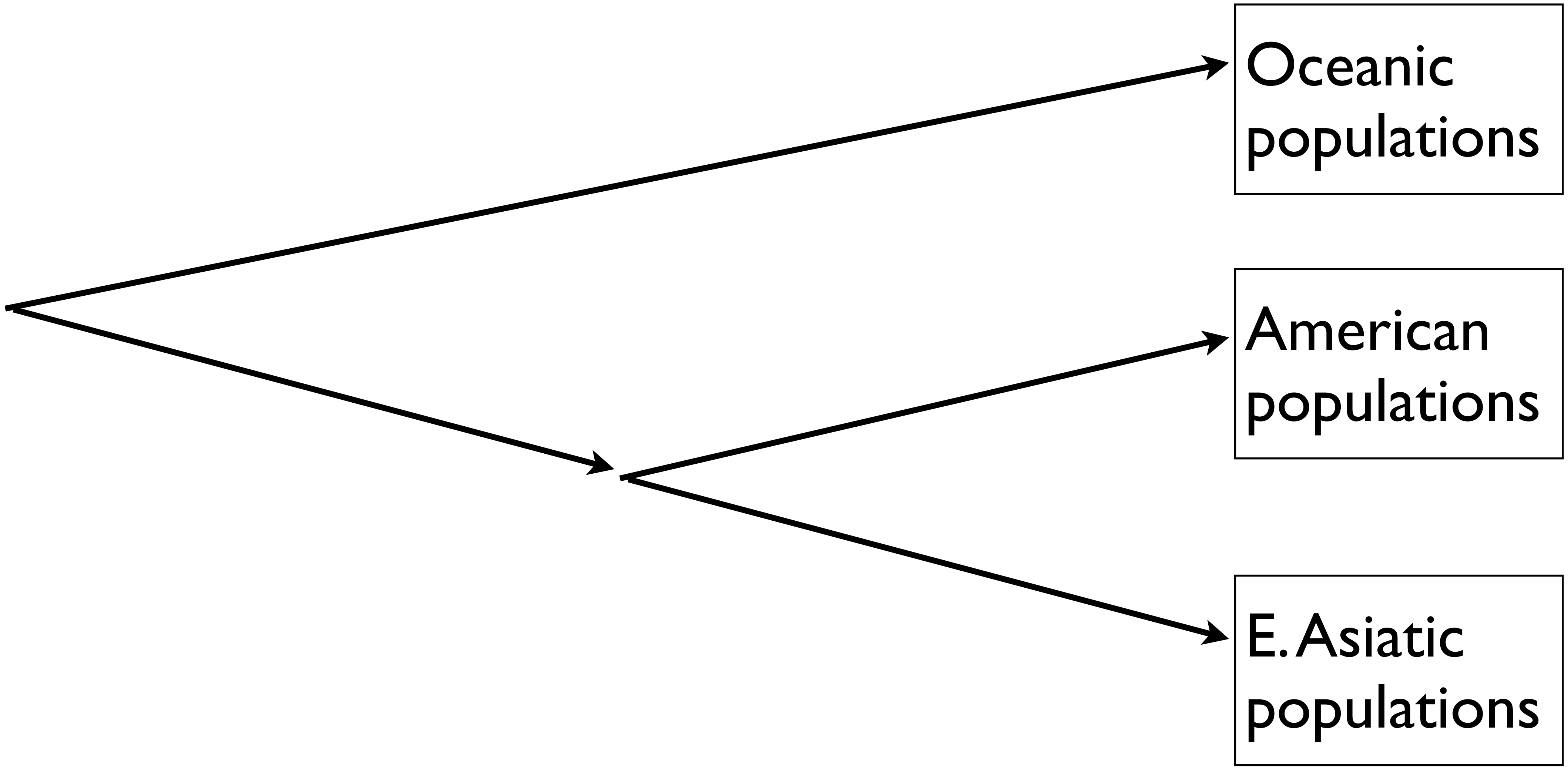

Observed
Oceanic
frequency
$P_{Oc.}(A)$
Same frequency
Same frequency
initially
Oceanic
populations
American
populations
Observed
American
frequency
$P_{Am.}(A)$
E. Asiatic
populations
Observed
E.Asiatic
frequency
$P_{E.A.}(A)$
All the ancestral frequencies
are unknown: integrate over
them



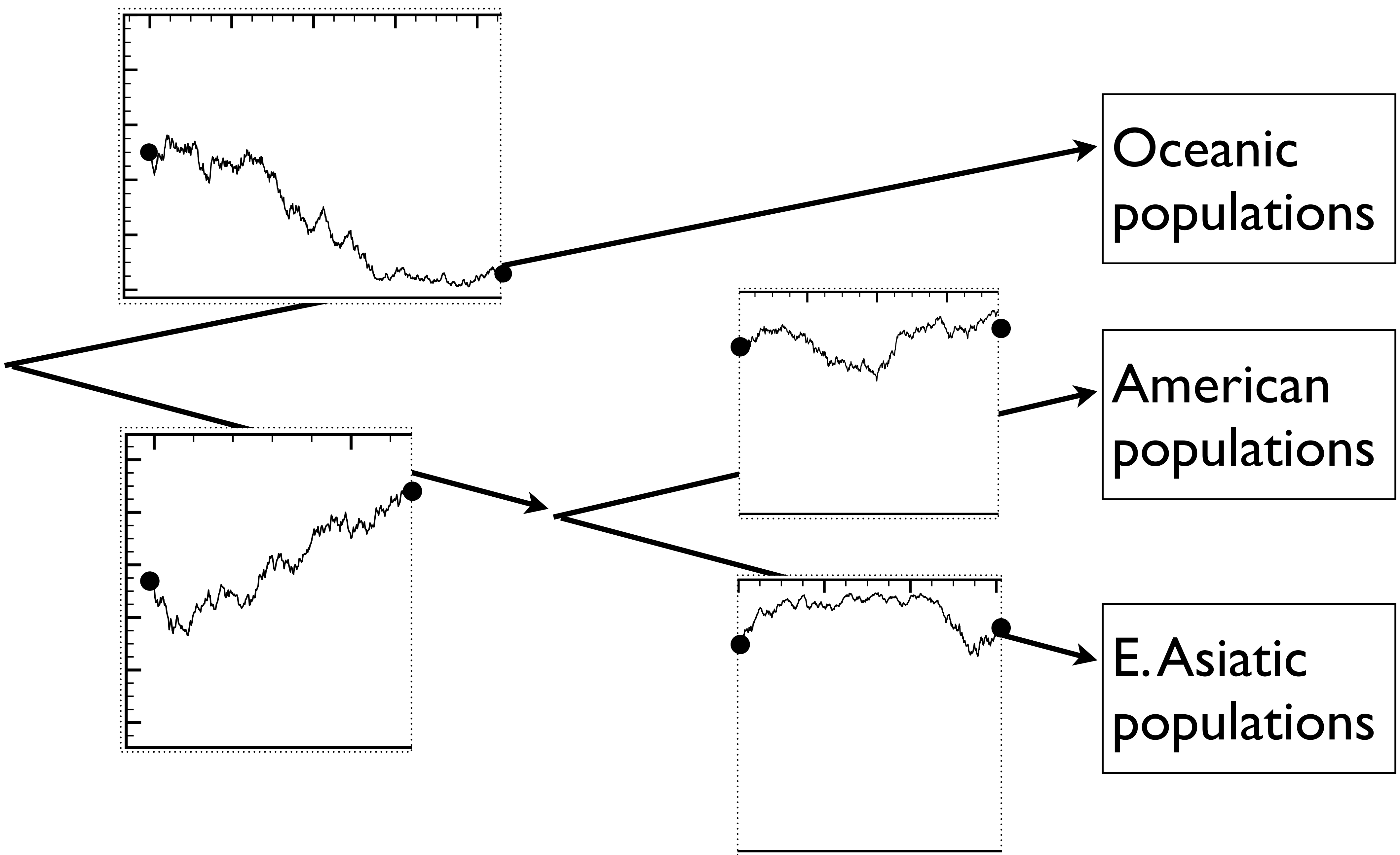
Same frequency
Observed Oceanic frequency $P_{Oc.}(A)$
Same frequency initially
Oceanic populations
American populations
Observed American frequency $P_{Am.}(A)$
E. Asiatic populations
Observed E.Asiatic frequency $P_{E.A.}(A)$
All the ancestral frequencies are unknown: integrate over them
Doing this for each SNP gives us P(Data | H1)


Oceanic populations
American populations
E. Asiatic populations

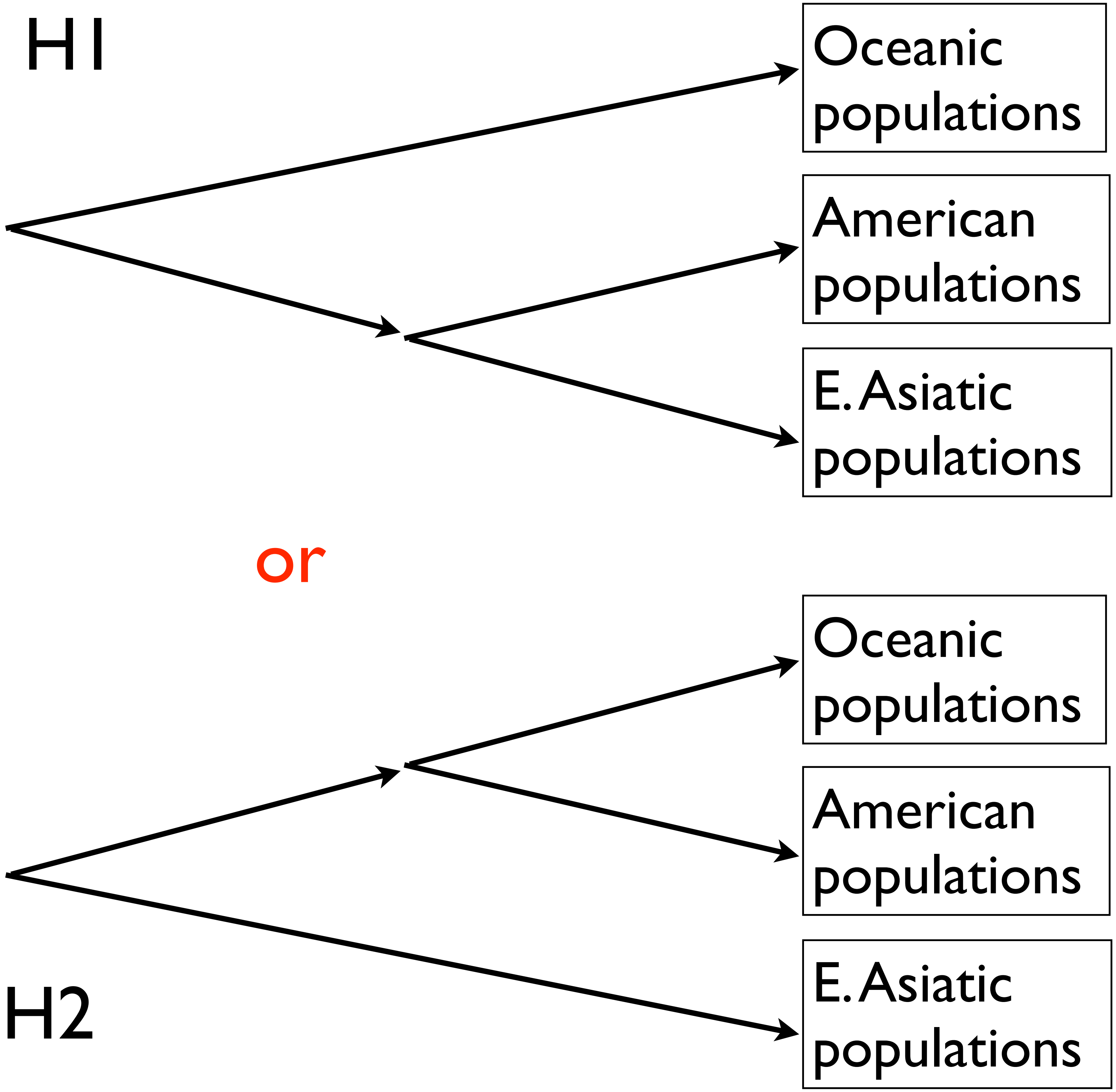
Doing the same thing, but with the other tree gives us P(Data | H2)



Doing the same thing, but with the other tree gives us P(Data | H2)



Simplified example



$$\frac{P(\text{Data} \mid H1)}{P(\text{Data} \mid H2)}$$

Questions:

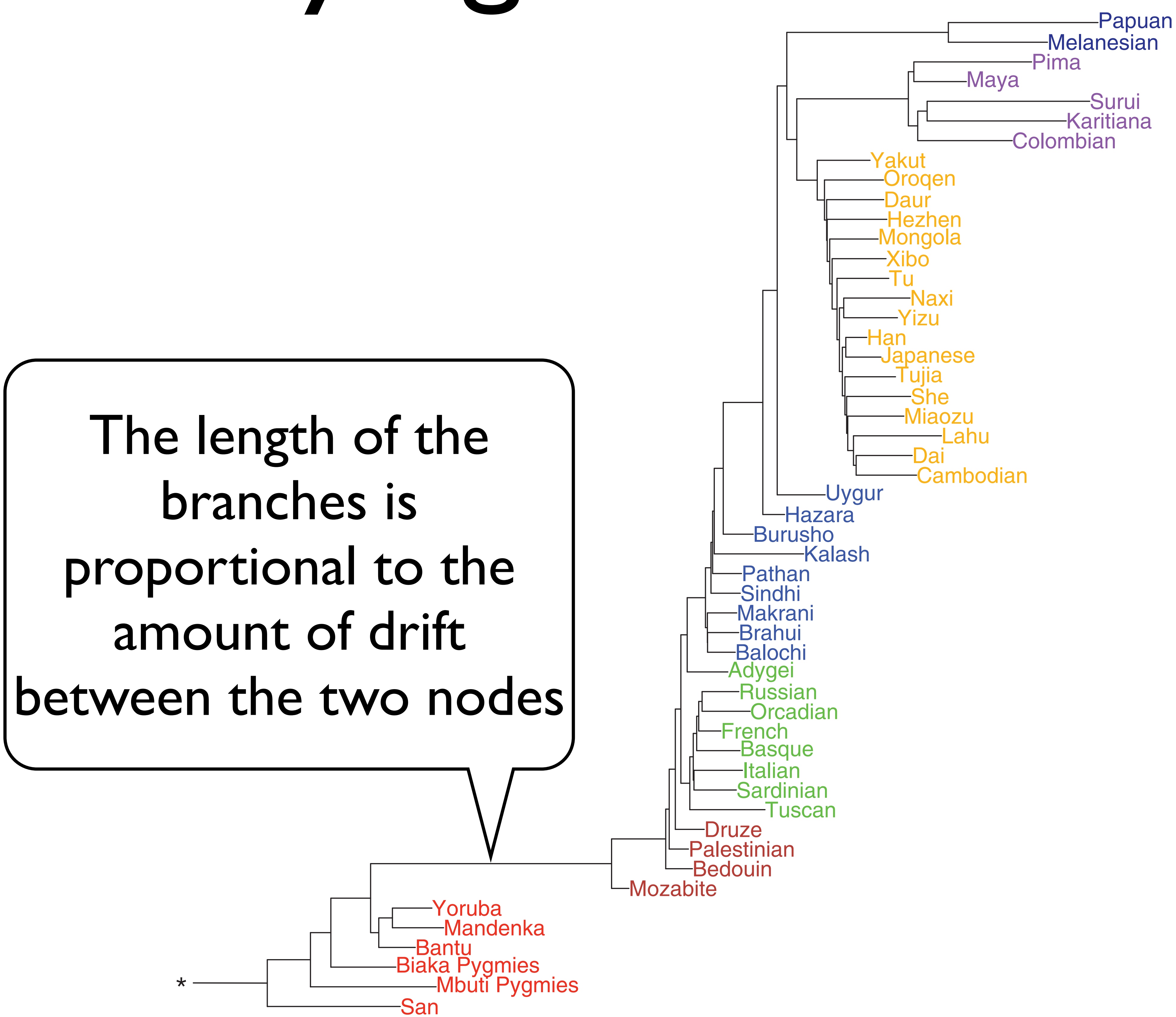
<u>What is the data?</u>

Allele frequencies for each population

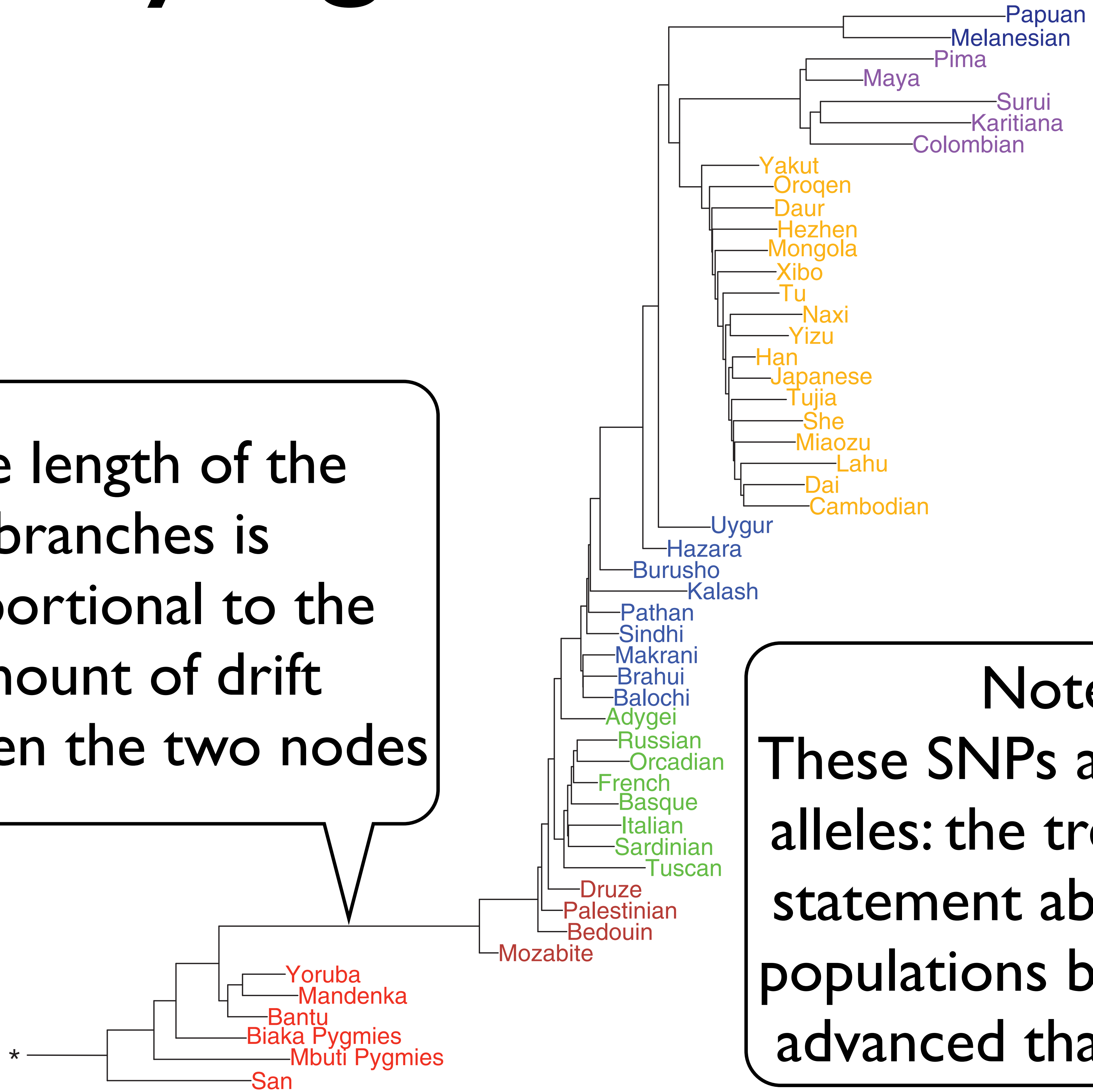
<u>What is the model, i.e. what is P?</u>

Wright-Fisher model (Brownian motion approximation)

Phylogenetic tree



Phylogenetic tree



The length of the branches is proportional to the amount of drift between the two nodes

Note:
These SNPs are neutral alleles: the tree is *not* a statement about some populations being more advanced than others!