Mismatch Distributions and Population Growth

Alan R. Rogers

February 10, 2022

What is a mismatch distribution?

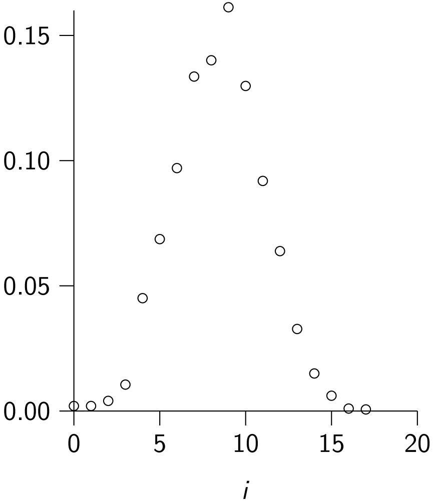
Count the number of site differences between each pair of sequences in a sample, and use the resulting counts to build a histogram. You end up with a “mismatch distribution.” The *i*th entry of the mismatch distribution is the number of pairs of sequences that differ by *i* sites.

Partial mtDNA sequences from Asia

```
 1 CATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGG...
 2 CATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGG...
 3 CATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGG...
 4 CATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGG...
 5 CATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGG...
 6 CATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGG...
 7 CATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGG...
 8 CATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGG...
 9 CATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGG...
10 CATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGG...
11 CATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGG...
12 CATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGG...
13 CATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGG...
14 CATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGG...
15 CATTAGCACCCAAAGCTAAGATTCTAATTTAAACTATTCTCTGTTCTTTCATGG...
..............................................................
```

Mismatch distribution for Asian data

i	*n*	*i*	*n*
0	5	10	379
1	5	11	268
2	11	12	186
3	30	13	95
4	131	14	43
5	200	15	17
6	283	16	2
7	390	17	1
8	409		
9	471		

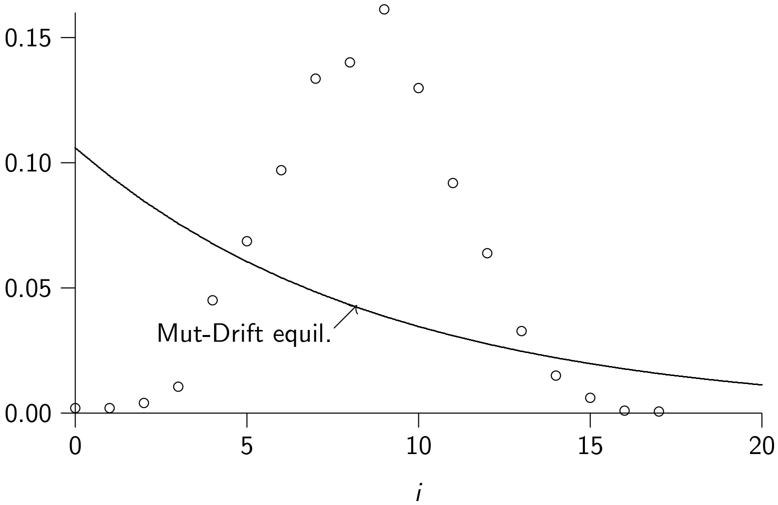


At mutation-drift equilibrium, a random pair of sequences differs by i sites with probability

$$F_i = \left(\frac{1}{\theta+1}\right)\left(\frac{\theta}{\theta+1}\right)^i, \qquad (i = 0, 1, 2, \ldots) \tag{1}$$

(Watterson, 1975)

mtDNA mismatch distribution doesn't fit equilibrium model



Why does the stationary neutral model fit human data so poorly?

There are several hypotheses to consider:

1. Sampling error. (Important because the pairs of genes in our sample are correlated.)
2. Selection.
3. Failure of infinite sites hypothesis.
4. Non-random mating.
5. Variation in population size.

Work has been done on all of these possibilities, but I will only try to tell you about the last one.

Coalescent theory in a population of varying size

At any given time, t, the hazard of a coalescent event is

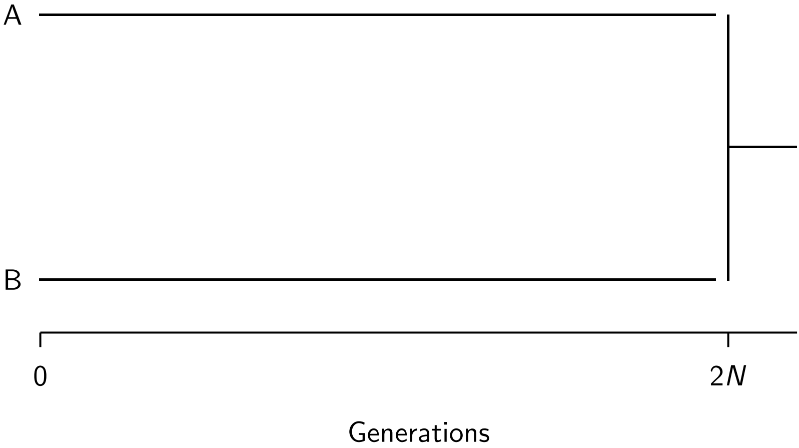
$$h_i(t) = \frac{i(i-1)}{4N(t)}$$

But $N(t)$ is no longer constant.

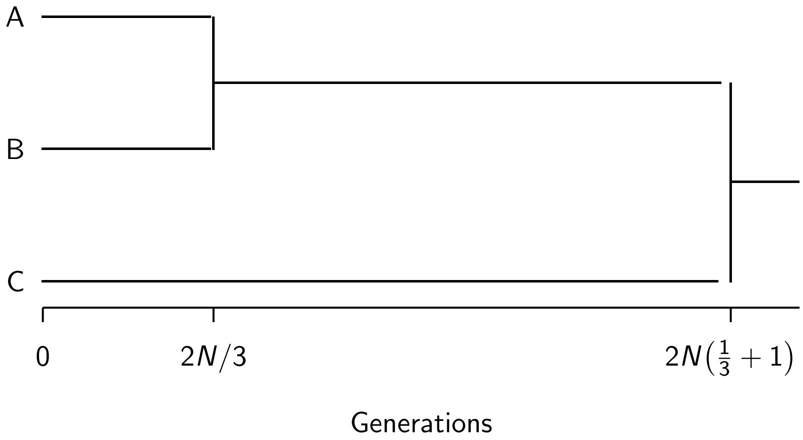
$$E[t_i] \neq 1/h_i$$

We need computer simulations.

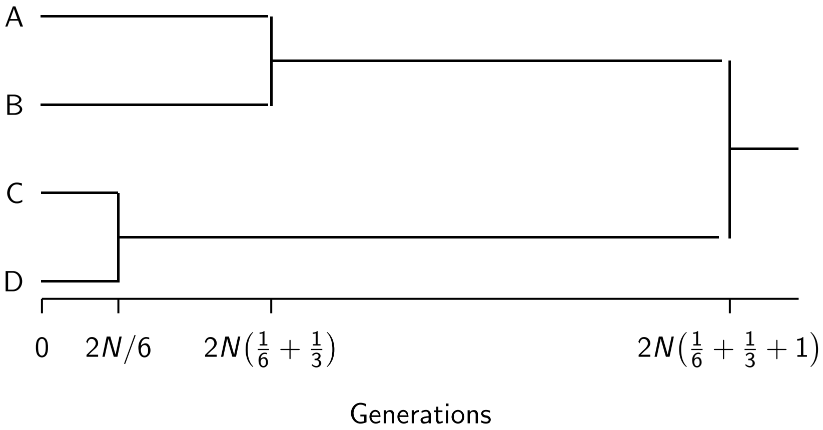
Expected genealogy of 2 genes



Expected genealogy of 3 genes



Expected genealogy of 4 genes


A
B
C
D
0
$2N/6$
$2N\left(\frac{1}{6}+\frac{1}{3}\right)$
$2N\left(\frac{1}{6}+\frac{1}{3}+1\right)$
Generations

Principles

The expected length of a coalescent interval is long

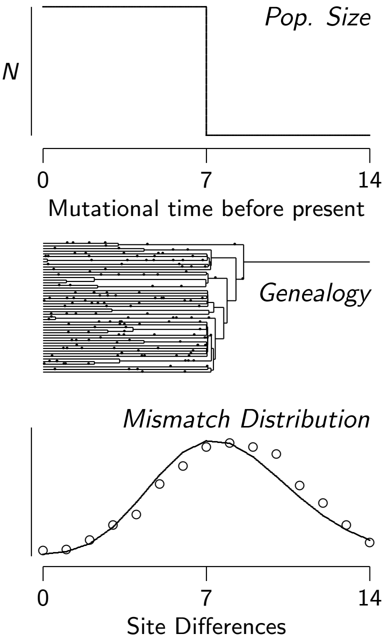
- in large populations
- if there are only a few lineages

Principles

The expected length of a coalescent interval is long

- in large populations
- if there are only a few lineages

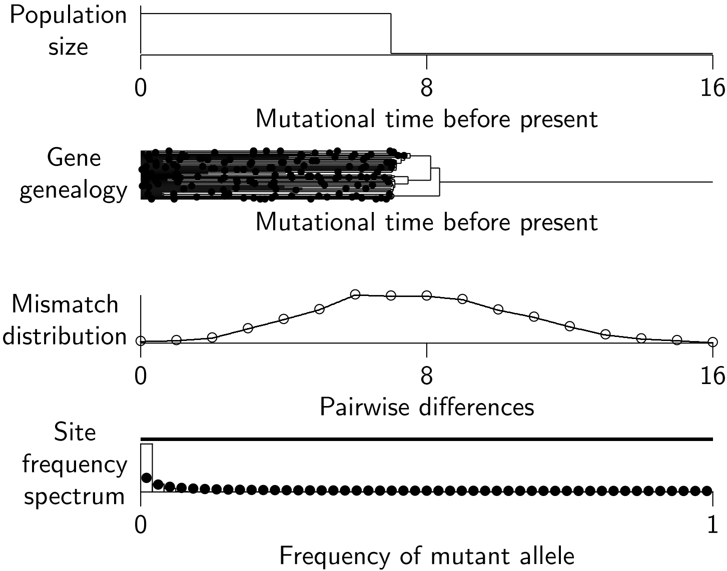
What if the population changes in size?



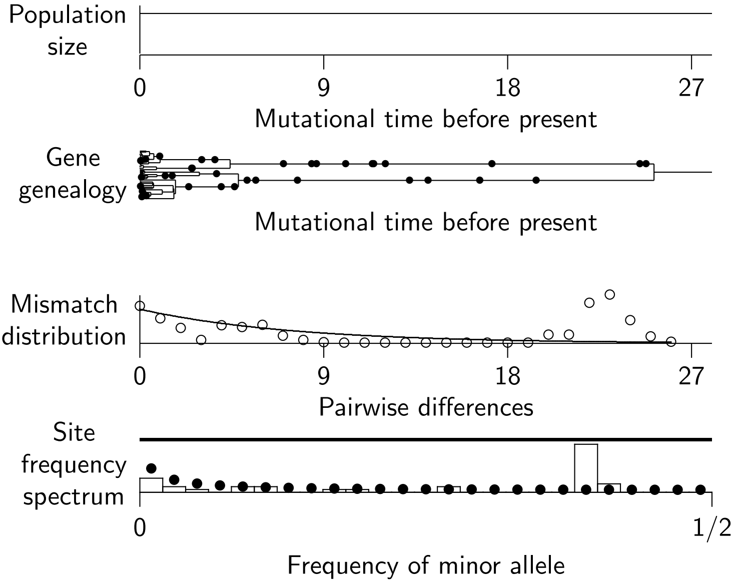
Effect of a population explosion

Middle: genealogy of 50 individuals; dots are mutations.

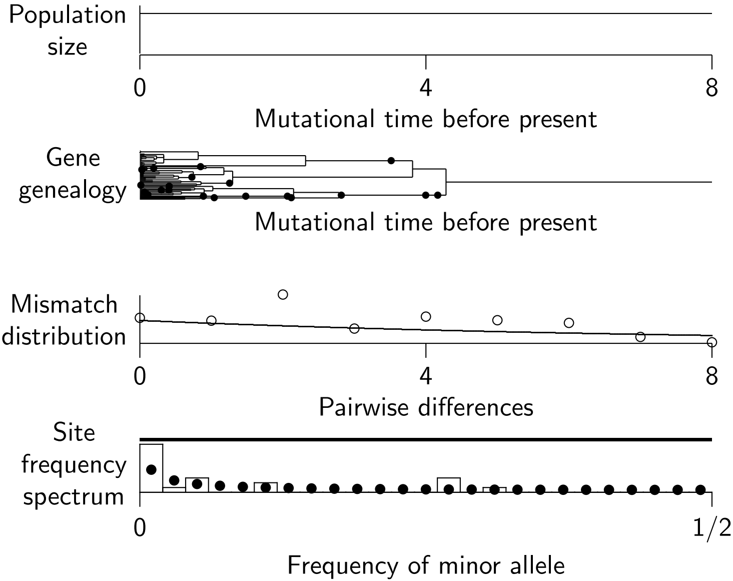
Bottom: ○ = simulated data, line = theory.


Population size
0
8
16
Mutational time before present
Gene genealogy
Mutational time before present
Mismatch distribution
0
8
16
Pairwise differences
Site frequency spectrum
0
1
Frequency of mutant allele

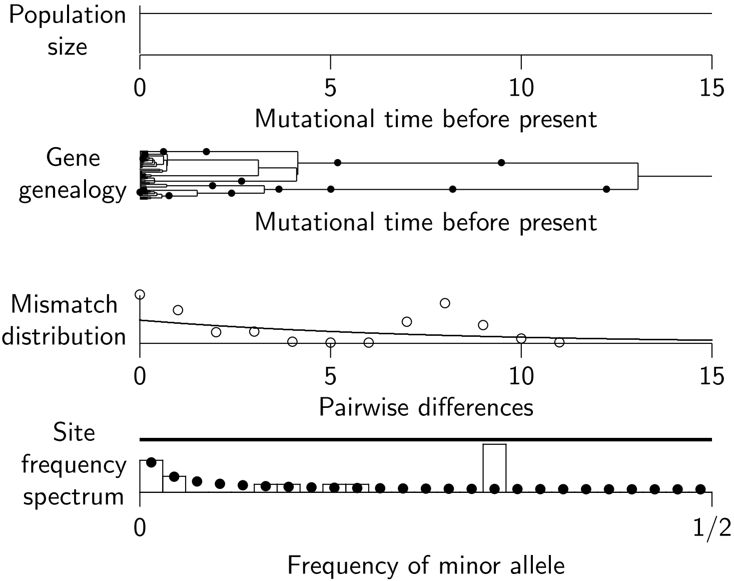
Simulations of stationary populations



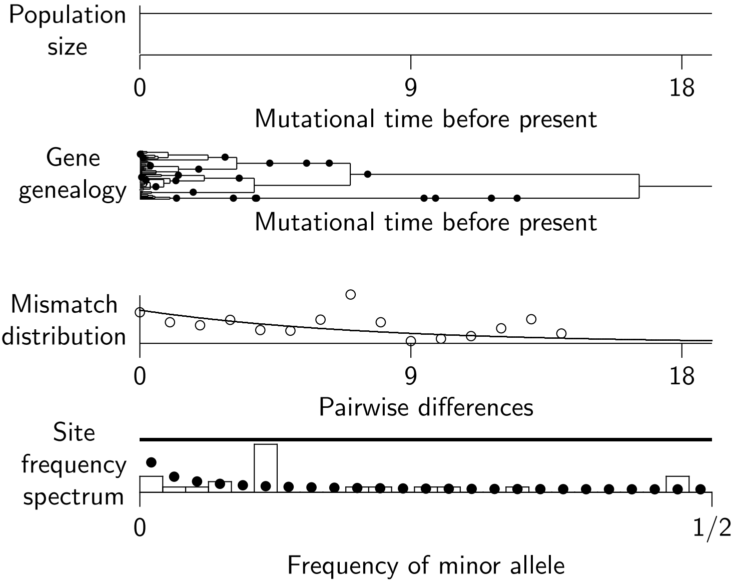
An equilibrium population with $\theta = 7$.



An equilibrium population with $\theta = 7$.

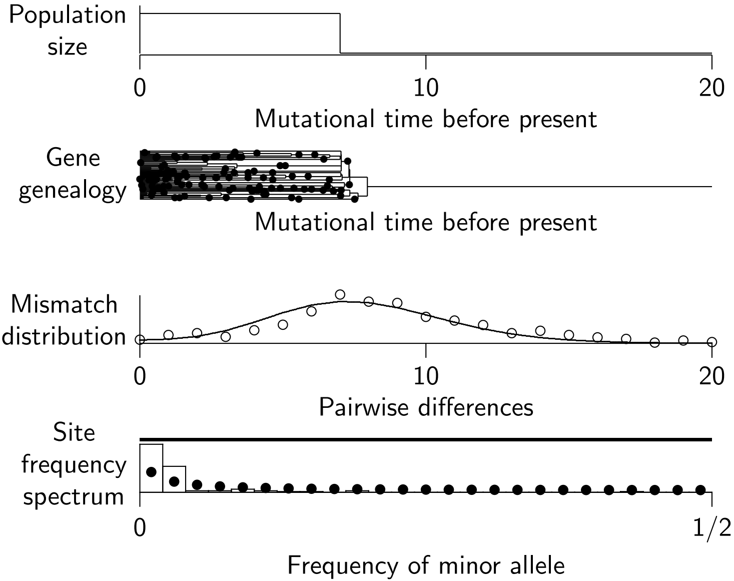


An equilibrium population with $\theta = 7$.

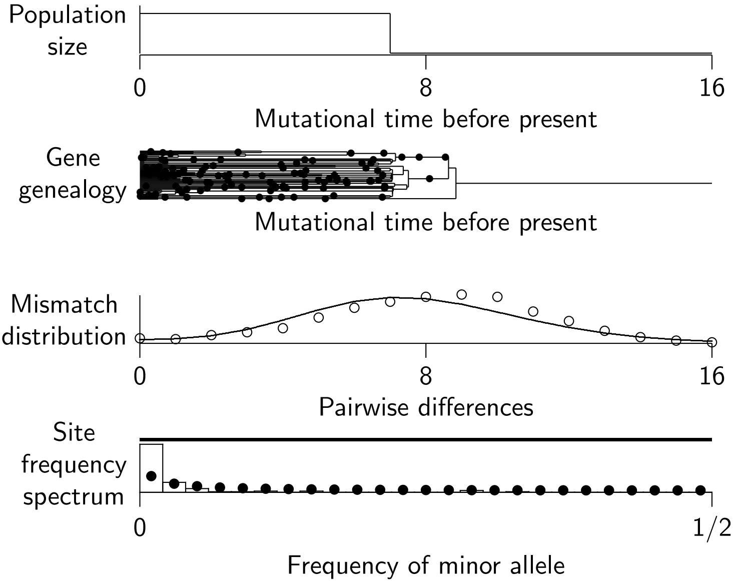


An equilibrium population with $\theta = 7$.

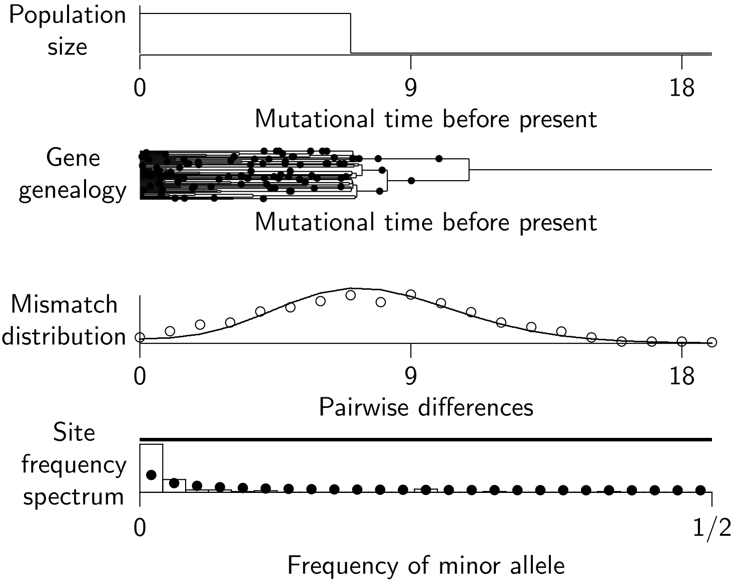
Simulations of expanded populations



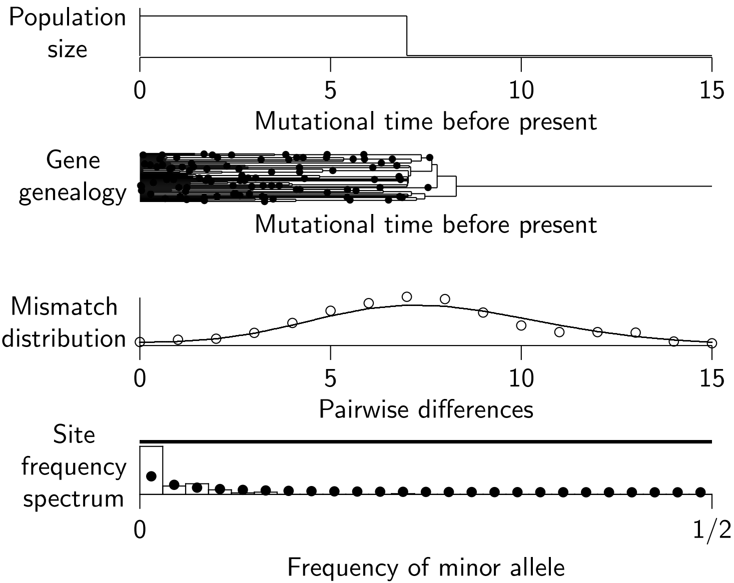
A coalescent simulation with growth: $\theta_0 = 1$, $\theta_1 = 100$, $\tau = 7$.



A coalescent simulation with growth: $\theta_0 = 1$, $\theta_1 = 100$, $\tau = 7$.



A coalescent simulation with growth: $\theta_0 = 1$, $\theta_1 = 100$, $\tau = 7$.



A coalescent simulation with growth: $\theta_0 = 1$, $\theta_1 = 100$, $\tau = 7$.

Model of sudden change in population size

Time variable: $\tau = 2\mu t$, where μ is the mutation rate and t is time in generations.

Population size changes suddenly at time 0. $\theta = 4N\mu$ measures population size after this change.

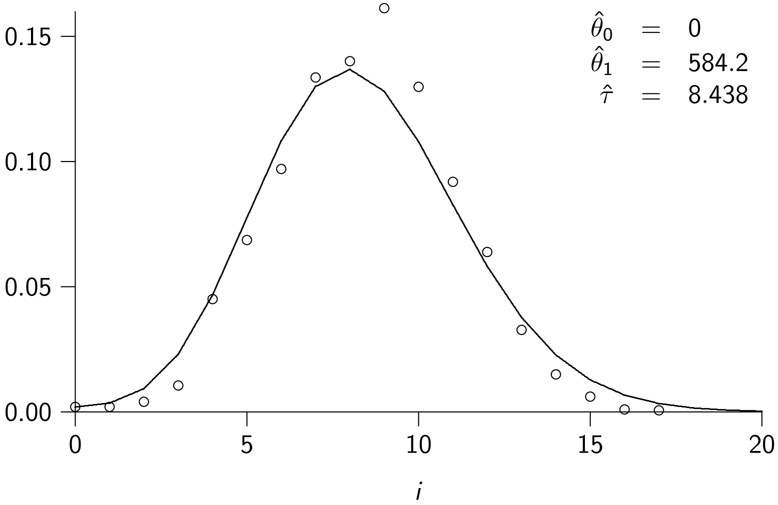
At time 0, $F_i(0)$ is probability that a random pair of sequences differ by i sites. At time τ, this probability becomes $F_i(\tau)$. Converges toward an equilibrium at which $F_i(\tau) = \hat{F}_i$.

At time τ,

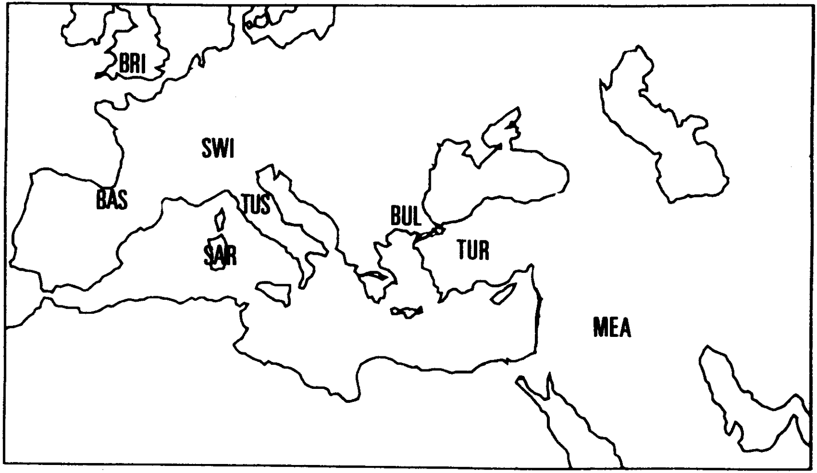
$$F_i(\tau) = \hat{F}_i + e^{-\tau(1+1/\theta)} \sum_{j=0}^{i} \frac{\tau^j}{j!} \left(F_{i-j}(0) - \hat{F}_{i-j}\right).$$

(Li, 1977; Rogers & Harpending, 1992)

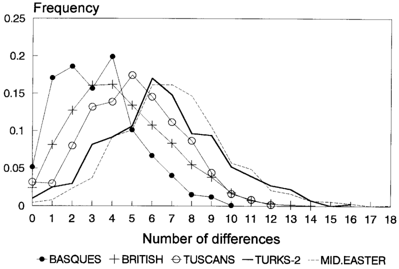
Model of sudden growth fit to Asian data

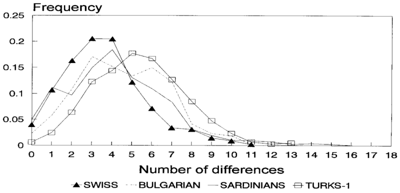


Comas et al (1997) studied European mismatch distributions



Mismatch distributions suggest expansion across Europe





Mid-East and Turkey: early expansions.

British, Basques: late expansions

Paleolithic or Neolithic?

Comas et al thought Paleolithic but may have been misled by mtDNA clock.