Mismatch Distributions and Population Growth

Alan R. Rogers

October 2, 2015

What is a mismatch distribution?

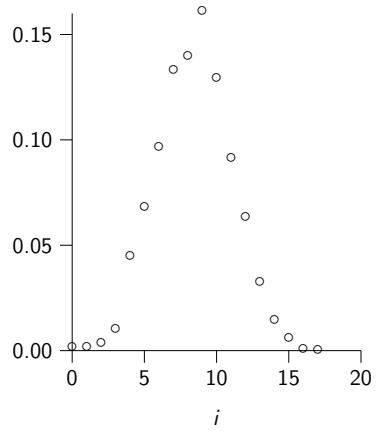
Count the number of site differences between each pair of sequences in a sample, and use the resulting counts to build a histogram. You end up with a "mismatch distribution." The ith entry of the mismatch distribution is the number of pairs of sequences that differ by i sites.

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

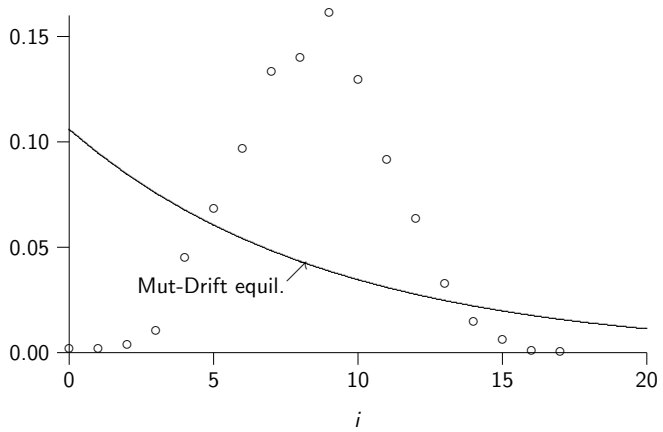
i	n	i	n
0	5	10	379
1	5	11	268
2	11	12	186
3	30	13	95
4	131	14	43
5	200	15	17
6	283	16	2
7	390	17	1
8	409		
9	471		



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

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

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.

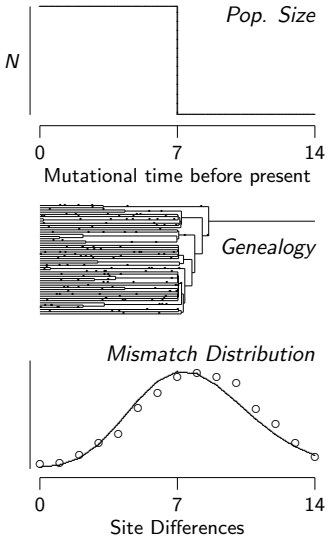
$$E[t_i] \neq 1/h_i$$

We need computer simulations.

Principles

Coalescent intervals tend to be long in those parts of the tree where

- there are only a few lines of descent
- the population size is large

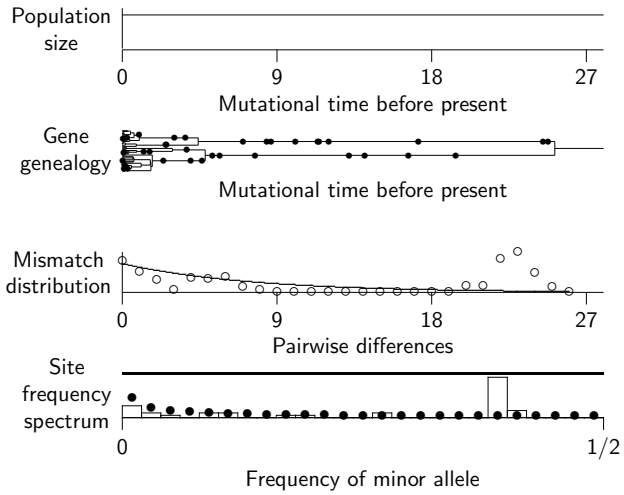


Effect of a population explosion

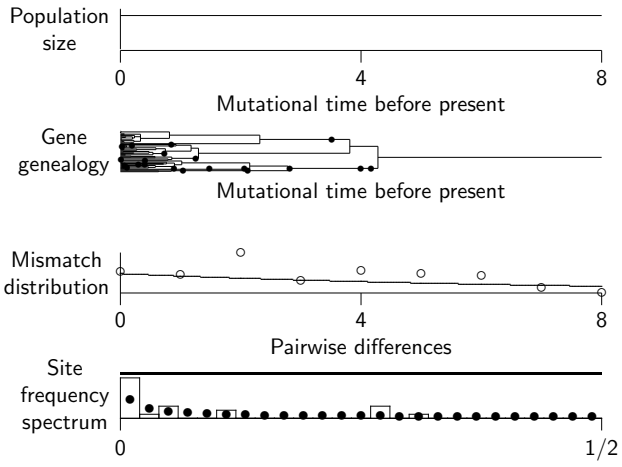
Middle: genealogy of 50 individuals; dots are mutations.

Bottom: ○ = simulated data, line = theory.

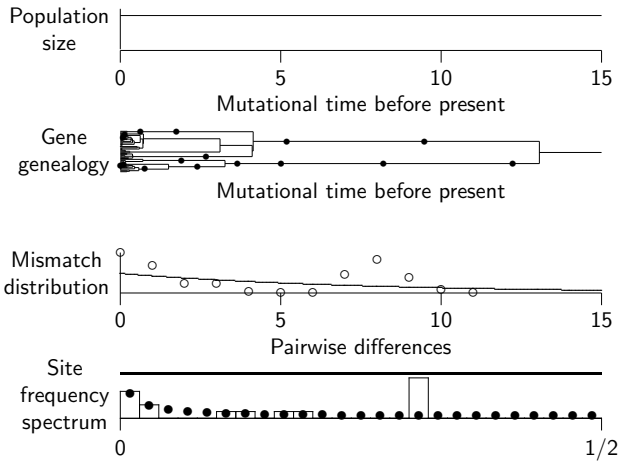
Simulations of stationary populations



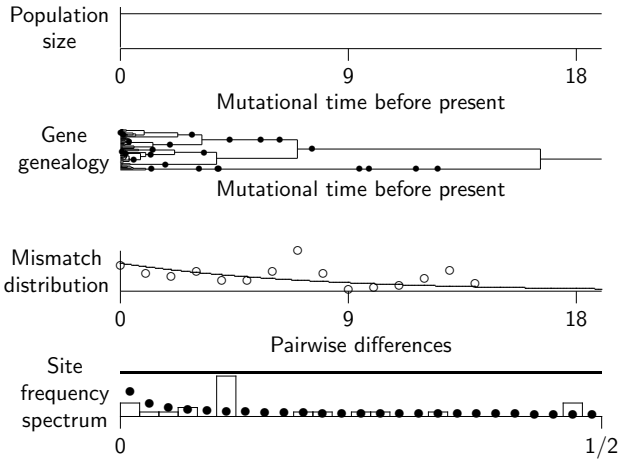
An equilibrium population with $\theta = 7$



An equilibrium population with $\theta = 7$

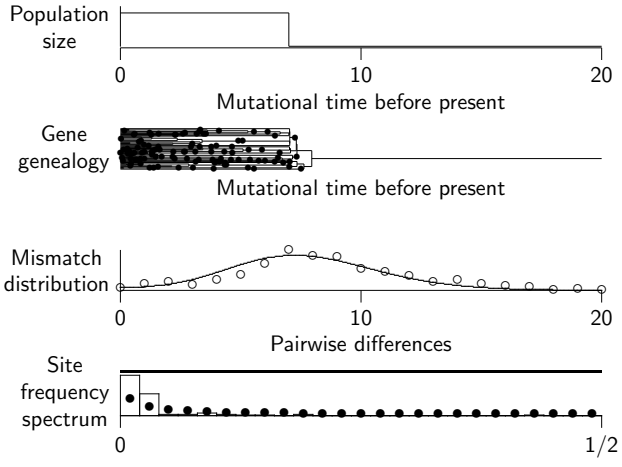


An equilibrium population with $\theta = 7$

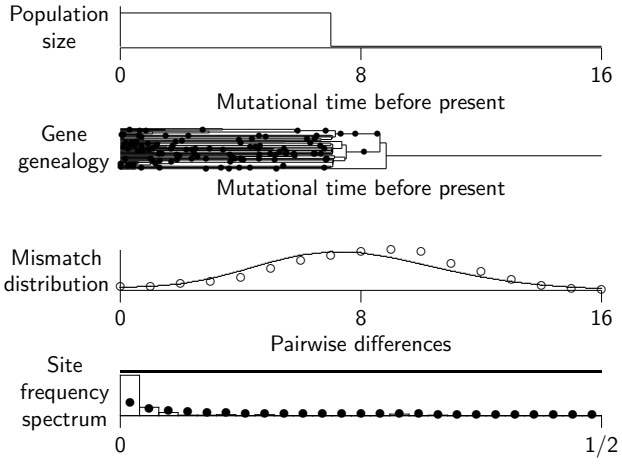


An equilibrium population with $\theta = 7$

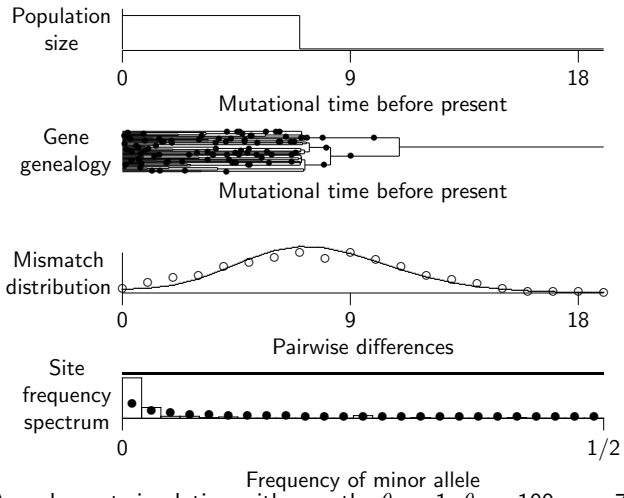
Simulations of expanded populations



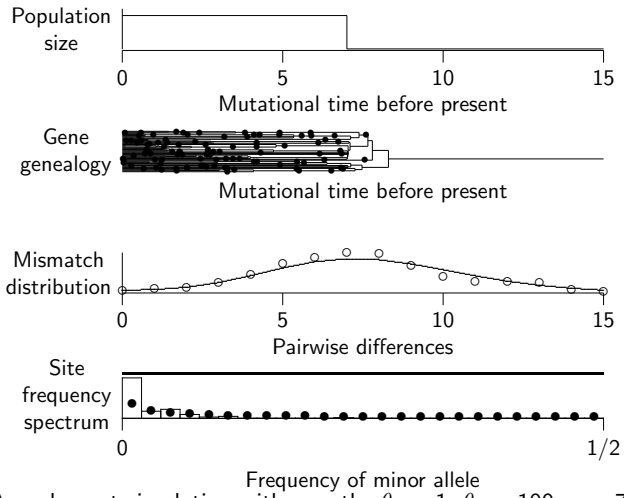
A coalescent simulation with growth: $\theta_0 = 1$, $\theta_1 = 100$, $\tau = 7$



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 growth fit to Asian data