Summarizing and using genetic variation

Mark Reppell

July 21, 2015

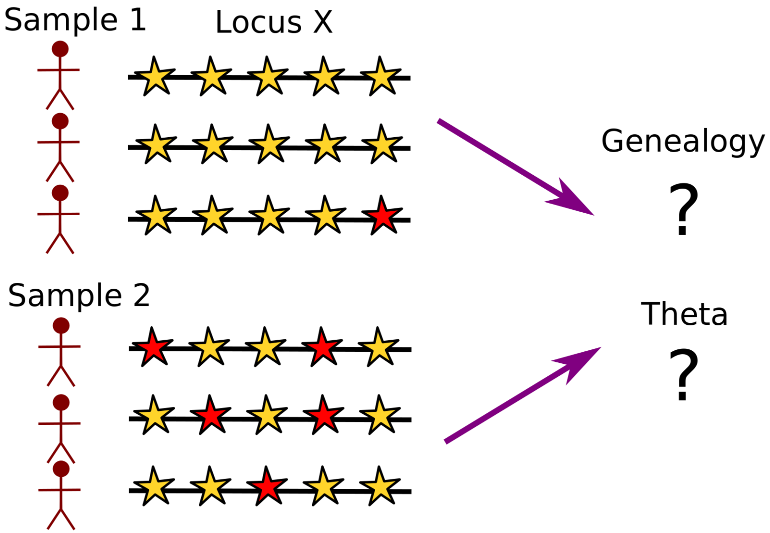
What can we learn about a population from observed genetic sequences?

Unfortunately, in real populations we can't flip the -T flag and work with the genealogies that underlie a sample of chromosomes

We **can** observe the quantity and counts of genetic variants in a sample

How can we summarize and use this genetic variation to learn about the underlying population?

Using the quantity and pattern of variants



Outline

- The site frequency spectrum
- Estimating θ
- Tajima's D and detecting departure from a neutral sequence model

The expected number of segregating sites

The times between coalescent events have distribution $T_k \sim Exp(\binom{k}{2})$, with $k \in (2, n)$

$$E(T_{total}) = \sum_{k=2}^{n} kE(T_k) = \sum_{k=2}^{n} k\frac{2}{k(k-1)} = 2\sum_{k=1}^{n-1}\frac{1}{k}$$

So, given $E(\text{mutations}|\text{time}) = \frac{\theta t}{2}$

$$E(\text{Segregating sites}) = \frac{\theta}{2}\left(2\sum_{k=1}^{n-1}\frac{1}{k}\right) = \theta\sum_{k=1}^{n-1}\frac{1}{k}$$

The site frequency spectrum

The frequency spectrum counts the number of derived alleles observed in i members of a sample size n, for $i \in (1, n-1)$

```
-C-T-A-T-A-T-C-G-C
-C-T-A-T-A-T-C-G-C
-T-T-A-T-A-T-C-G-C
-T-T-A-T-A-T-C-G-C
-C-G-A-T-A-A-C-T-C
-T-T-A-T-A-A-C-T-C
-C-T-C-T-A-T-C-T-C
-C-T-A-T-A-T-G-T-A
-C-T-A-G-A-T-C-T-A
-C-T-A-G-C-T-C-T-A
```

$$\xi_1 = 3$$
$$\xi_2 = 2$$
$$\xi_3 = 1$$
$$\xi_4 = 1$$
$$\xi_5 = 0$$
$$\xi_6 = 1$$
$$\xi_7 = 0$$
$$\xi_8 = 1$$
$$\xi_9 = 0$$

$$S = \sum_{i=1}^{n-1} \xi_i$$

The folded site frequency spectrum

The derived allele is not always known, so the folded spectrum, which counts the number of minor alleles observed, is used instead of the full frequency spectrum

```
-C-T-A-T-A-T-C-G-C
-C-T-A-T-A-T-C-G-C
-T-T-A-T-A-T-C-G-C
-T-T-A-T-A-T-C-G-C
-C-G-A-T-A-A-C-T-C
-T-T-A-T-A-A-C-T-C
-C-T-C-T-A-T-C-T-C
-C-T-A-T-A-T-G-T-A
-C-T-A-G-A-T-C-T-A
-C-T-A-G-C-T-C-T-A
```

$$\eta_i = \begin{cases} \xi_i + \xi_{n-i} & \text{if } i \neq n-i \\ \xi_i & \text{if } i = n-i \end{cases}$$

$$\eta_1 = 4$$

$$\eta_2 = 2$$

$$\eta_3 = 2$$

$$\eta_4 = 1$$

$$\eta_5 = 0$$

The expected frequency spectrum

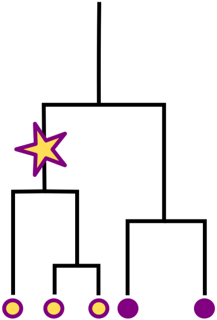
For a variant to occur *i* times in a sample, a mutation has to occur along a branch with exactly *i* descendants.

Such a branch is said to have size *i*, or to subtend *i* external nodes.

Not all tree have branches of all sizes

If τ_i is the sum of lengths (scaled) of branches of size *i*

$$E(\xi_i) = \frac{\theta}{2} E(\tau_i)$$



Expected SFS entries

$E(\tau_i)$ can be written as

$$E(\tau_i) = \sum_{k=2}^{n} k p_{n,k}(i) E(T_k)$$

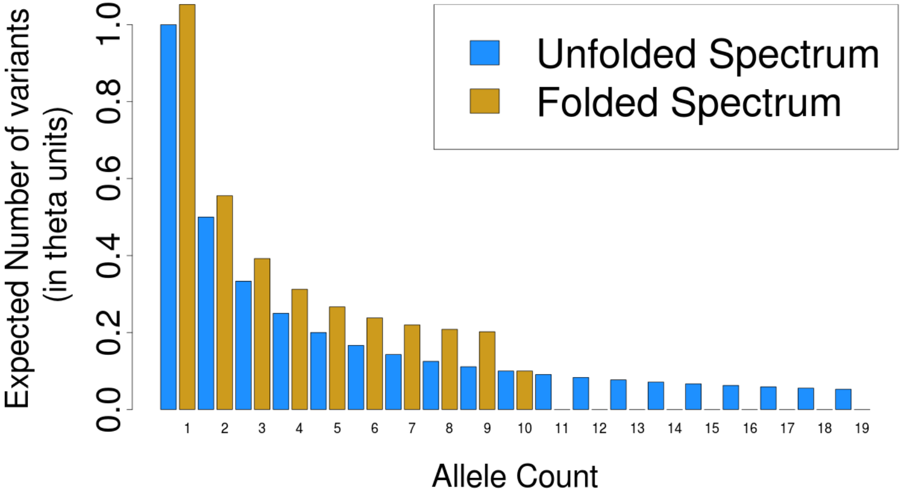
where $p_{n,k}(i)$ is the probability in a sample of size n that a branch while k branches remain has size i. Griffiths (1998) used an urn model to show

$$p_{n,k}(i) = \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \quad \text{and} \quad E(\tau_i) = \sum_{k=2}^{n} k \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \frac{2}{k(k-1)} = \frac{2}{i}$$

So the expected number of sites with size i is:

$$E(\xi_i) = \frac{\theta}{2} \times \frac{2}{i} = \frac{\theta}{i}$$

The frequency spectrum for n=20



Watterson estimator

If S is the total number of segregating sites observed in a sample with size n, and

$$E(S) = \theta \sum_{k=1}^{n-1} \frac{1}{k}$$

Then a simple estimator of θ is

$$\hat{\theta}_W = \frac{S}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

This is known as the Watterson estimator, and is unbiased for θ

Mean pairwise difference

Introduced by Nei and Li (1979), mean pairwise sequence difference is commonly written as π

```
C-T-A-T-A-T-C-G-C
T-T-A-T-A-T-C-G-C
C-G-A-T-A-A-C-T-C
T-T-A-T-A-A-C-T-C
```

If $k_{i,j}$ is the number of difference between sequences i and j

$$\pi = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k_{i,j}$$

	Sample2	Sample3	Sample4
Sample1	1	3	3
Sample2	-	4	2
Sample3	-	-	2

So, in the example at left

$$\pi = \frac{1}{6}(1+2+2+3+3+4) = 2.5$$

Mean pairwise difference and heterozygosity

π can be considered on a per site basis

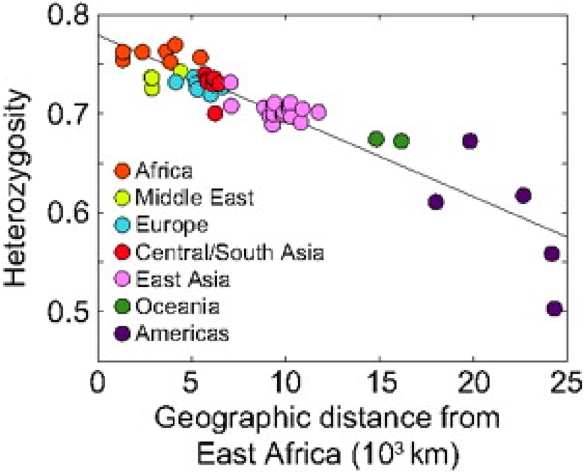
$$\pi = \sum_{i=1}^{m} \pi_i$$

At dimorphic site i, if allele A is observed k times

$$\pi_i = \frac{k(n-k)}{\binom{n}{2}} = 2 \cdot \frac{k}{n} \cdot \frac{n-k}{n-1} \approx 2\hat{p}(1-\hat{p})$$

π can be considered the sum of heterozygosities at polymorphic sites in a sequence.

Heterozygosity in human populations



DeGiorgio *et al.* (2009)

Estimating θ using π

$$E(\pi) = E\left(\binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{i,j}\right) = \binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E(k_{i,j})$$

From its original definition, $E(k_{i,j}) = \theta$

$$\binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \theta = \frac{\theta}{\binom{n}{2}}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} 1 = \frac{\theta}{\binom{n}{2}}\binom{n}{2} = \theta$$

Therefore, π is also an unbiased estimator of θ

Tajima's D

$\hat{\theta}_W$ and π are both estimators of θ, and if model assumptions are correct they should give the same value

Tajima's D (1989) is a statistic to test $\hat{\theta}_W = \pi$

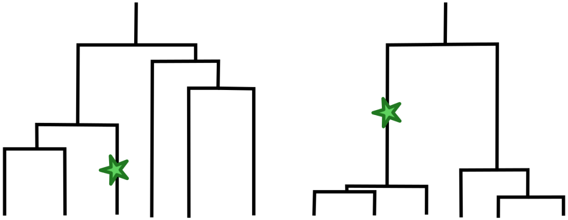
$$D = \frac{\pi - \hat{\theta}_W}{\sqrt{\hat{Var}(\pi - \hat{\theta}_W)}} = \frac{\pi - \frac{S}{a_1}}{\sqrt{\hat{Var}(\pi - \frac{S}{a_1})}} \text{ with } a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$$

$$Var(\pi - S/a_1) = Var(\pi) - 2Cov(\pi, S)/a_1 + Var(S)/a_1^2$$

D has an expectation of 0 and variance ≈ 1

The behavior of $\hat{\theta}_W$ and π

Deviations from the neutral model alter the structure of sample genealogies



Both cases add 1 mutation $\rightarrow \hat{\theta}_W + \frac{1}{\sum_{i=1}^{5} \frac{1}{i}} = \hat{\theta}_W + 0.44$

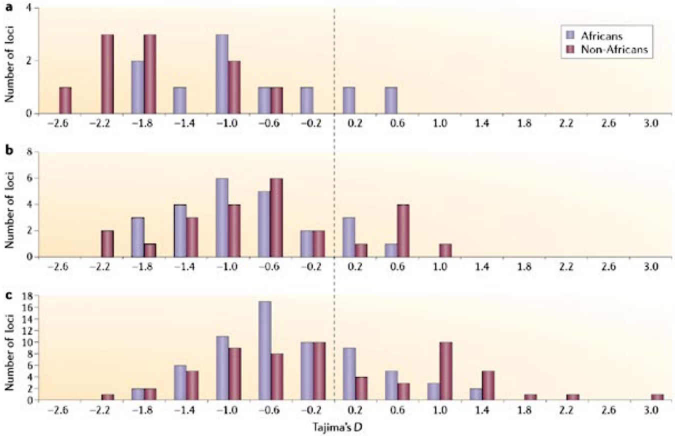
The left adds 5 new differences $\rightarrow \pi + \frac{5}{\binom{6}{2}} = \pi + \frac{1}{3}$

The second adds 9 $\rightarrow \pi + \frac{9}{\binom{6}{2}} = \pi + \frac{3}{5}$

Interpretation of Tajima's D

$D_T < -2$	$-2 < D_T < 2$	$D_T > 2$
$\theta_\pi < \theta_W$	$\theta_\pi = \theta_W$	$\theta_\pi > \theta_W$
Long external branches	Standard tree	Long internal branches
Selective sweep	Neutral model	Population subdivision
Population expansion	Low power	Shrinking population

Tajima's D in human populations



Nature Reviews | Genetics

a) mtDNA and Y-Chromosome **b)** X-Chromosome **c)** Autosomes

Garrigan and Hammer (2006)

Lab 2

- Compiling and running the program *sample_stats*
- Plotting summary statistics in R
- Impact of population size and mutation rate on sequence characteristics