Calculating genetic diversity

Heterozygosity, Tajima's π, and Watterson's θ_W

George P. Tiley

2026-08-25

Table of contents

The Data

Four diploid individuals each from two populations were sequenced at the same 20 base pair (bp) locus. Every individual contributes two sequences, which are indicated by an `a` and `b` haplotype. Since organisms are diploid, there are eight sequences per population and sixteen total. One individual failed partially during sequencing.

```
          0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 2
          1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0
          ---------------------------------------
P1_1A     A C G T A C G T A C G T A C G T A C G T
P1_1b     A C G T A C G T A C G T A C G T A C G T
P1_2A     A C G T A C A T A C G T A C G T A C G T
P1_2b     A C G C A C A T A C G T A C G T A C G T
P1_3A     A C G C A C G T A C G T A C G T A C G T
P1_3b     A C G C A C G T A C G T A C G T A C G T
P1_4A     A C G T A C A T A C G T A C G T A C G T
P1_4b     A C G C A C A T A C G T A C G T A C G T

P2_1A     A C G T A C G T A C G T A C G T A C G T
P2_1b     A C G T A C G T A C A T A C G A A C G T
P2_2A     A C G T A C G T A C A T A C G T A C G T
P2_2b     A C G T A C G T A C A C A C G T A C C T
P2_3A     A C G T A C G T A C G T A C G A A C G T
P2_3b     A C G T A C G T A C A T A C G A A C G T
P2_4A     A C G T A C G T N N N N N N G T A C G T
P2_4b     A C G C A C G T N N N N N N G A A C G T
```

Finding the variable/segregating sites in each population separately

`N` means no data. Individual P2_4 failed at sites 9–14 on both chromosome copies.

Step 1 — Find the segregating sites

A site is **segregating** if more than one base shows up among the sequences you're looking at. Scan down each column. Do the two populations separately: a site can vary in one and be fixed in the other.

Population	Segregating sites	S
Pop 1	4, 7	2
Pop 2	4, 11, 12, 16, 19	5

The rule for missing data

Don't throw out the individual and don't throw out the site. Work **column by column**, and at each column use only the sequences that actually have a base there. Your sample size n therefore changes as you move along the locus: sites 9–14 have $n = 6$ sequences in Population 2, and every other site has $n = 8$.

This is simply counting the A_1, A_2 or A_a from the HW Example Slides.

Step 2 — Observed and expected heterozygosity

H_o is about *individuals*, so pair each **a** sequence with its **b** partner into a genotype first. H_e is about *allele frequencies*, so for that you just count bases down the column and ignore who they came from.

Every segregating site in this alignment has exactly two alleles. Call their frequencies p and q, with $p + q = 1$. p should be the major allele, but it is fine if you swap them — $2pq$ is the same either way.

$$H_o = \frac{\text{heterozygous individuals}}{\text{individuals scored}} \qquad H_e = 2pq$$

Count p and q over the *sequences* with data at that site — 8 at most sites, but 6 at sites 11 and 12 in Population 2.

If a site ever had three or more alleles, $2pq$ generalises to $1 - \sum p_i^2$, summing the squared frequency of every allele present. Nothing in this dataset needs that.

Fill out the information for each segregating site. If you write a computer program, you will have to store these pieces of information to memory.

Population 1

Site	Individuals	n seq.	Allele counts	Het. indiv.	H_o	$H_e = 2pq$
4	4	8	T=4, C=4	2	$\frac{2}{4} = 0.5$	$2(0.5 \times 0.5) = 0.5$
7	4	8	G=4, A=4	0	0	$2(0.5 \times 0.5) = 0.5$

Population 2

Site	Individuals	n seq.	Allele counts	Het. indiv.	H_o	$H_e = 2pq$
4	4	8	T=7, C=1	1	$\frac{1}{4} = 0.25$	$2(\frac{7}{8} \times \frac{1}{8}) = \frac{7}{32}$
11	3	6	A=4, G=2	2	$\frac{2}{3}$	$2(\frac{4}{6} \times \frac{2}{6}) = \frac{4}{9}$
12	3	6	T=5, C=1	1	$\frac{1}{3}$	$2(\frac{5}{6} \times \frac{1}{6}) = \frac{5}{18}$
16	4	8	T=4, A=4	2	$\frac{2}{4} = 0.5$	$2(0.5 \times 0.5) = 0.5$
19	4	8	G=7, C=1	1	$\frac{1}{4} = 0.25$	$2(\frac{7}{8} \times \frac{1}{8}) = \frac{7}{32}$

Site	Individuals	n seq.	Allele counts	Het. indiv.	H_o	$H_e = 2pq$

Now average across the whole locus. Add your per-site values and divide by **20**, not by the number of segregating sites — the monomorphic sites are real data and they count as zeroes.

Population	$\sum H_o$	mean H_o	$\sum H_e$	mean H_e
Pop 1	0.5	0.025	1	0.05
Pop 2	2	0.1	1.66	0.083

0.025. I almost messed up because I forgot the 18 invariable sites.

rounded. $\frac{239}{144} = 1.6597$

Step 3 — Nucleotide diversity, Tajima's π

π is the average number of differences between two randomly chosen sequences, per site. Forget individuals entirely here — all eight sequences in a population are independent observations.

$$\pi = \frac{1}{L} \sum_{\text{sites}} \frac{n}{n-1} \left(1 - \sum_i p_i^2\right)$$

$L = 20$ sites. The sum $\sum_i p_i^2$ runs over the alleles present *at that site* — and since every site here is biallelic, $1 - \sum_i p_i^2$ is just $2pq$, the quantity you already worked out as H_e in Step 2. Carry that column straight across.

So you do not need to calculate again.

n is the number of sequences with data at the site: 8 at most sites, but 6 at sites 11 and 12 in Population 2. The $n/(n-1)$ term corrects for the fact that you're working from a sample rather than knowing p and q outright ~~— more about this in the appendix.~~ Appendix was dropped. This helps correct for small samples and manages this missing data issue.

Population 1

Site	n	$1 - \sum_i p_i^2$	$\times\, n/(n-1)$	π_{site}
4	8	0.5	8/7	$4/7 \approx 0.57$
7	8	0.5	8/7	$4/7 \approx 0.57$

Population 2

Site	n	$1 - \sum_i p_i^2$	$\times\, n/(n-1)$	π_{site}
4	8	7/32	8/7	$1/4 = 0.25$
11	6	4/9	6/5	8/15
12	6	5/18	6/5	1/3
16	8	0.5	8/7	4/7
19	8	7/32	8/7	$1/4 = 0.25$

Sum the last column and divide by 20.

Pop 1: $\pi = \frac{8/7 \cdot 1/20}{} = \frac{4}{70} = 0.05714$

Pop 2: $\pi = \frac{1.9381}{20} = 0.09690$

I did not want to think about the fractions, so I just got the sum on my calculator.

Step 4 — Watterson's θ_W

π uses *how different* the sequences are from one another. Watterson's estimator ignores frequencies entirely and uses only *how many* sites vary, corrected for the fact that a bigger sample catches more rare variants.

$$\theta_W = \frac{1}{L} \sum_{\text{segregating sites}} \frac{1}{a_n} \qquad a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

Look up a_n by the number of sequences with data at that site.

n	2	3	4	5	6	7	8
a_n	1.0000	1.5000	1.8333	2.0833	2.2833	2.4500	2.5929

These are the only two values we will need.

The familiar shortcut is $\theta_W = S/a_n$, but that only works when every site has the same n. Population 2 doesn't, so each segregating site contributes its own $1/a_n$ and you add them up.

Population 1

Segregating site	n	a_n	$1/a_n$
4	8	2.5929	0.3857
7	8	2.5929	0.3857

$\sum_{i=1}^{S} \frac{1}{a_n} = 0.3857 + 0.3857 = 0.7714$

$\theta_W = \frac{1}{20} \times 0.7714 = 0.03857$

Population 2

Segregating site	n	a_n	$1/a_n$
4	8	2.5929	0.3857
11	6	2.2833	0.4380
12	6	2.2833	0.4380
16	8	2.5929	0.3857
19	8	2.5929	0.3857

$\sum_{i=1}^{S} \frac{1}{a_n} = 0.3857 + 0.4380 + 0.4380 + 0.3857 + 0.3857 = 2.0331$

Sum the last column and divide by 20.

Pop 1: $\theta_W =$ 0.03857 Pop 2: $\theta_W =$ 0.10165

$\theta_W = \frac{1}{20} \times 2.0331 = 0.10165$

Summary

Population	mean H_o	mean H_e	π	θ_W
Pop 1	0.025	0.05	0.057	0.039
Pop 2	0.1	0.083	0.097	0.102

Step 5 — Interpret

1. Your mean H_e and your π came out close but not equal — π is the larger of the two in both populations. Both are measuring the same thing: how often two randomly drawn sequences differ. Look at where the two calculations diverged. Which factor appears in one and not the other, and what is it doing there?

π assumes that $H_e = 2pq$ is biased low when the sample size is small. They should be equal as the sample size increases.

2. Given that, what is actually different between H_o and H_e? One of them would change if you shuffled which chromosome copies were packaged into which individuals, and the other would not. Which is which, and why does that matter?

H_o depends on chromosome pairing, but $H_e = 2pq$ does not. H_o measures the deviation from random expectations due to non-random mating.

3. π and θ_W both estimate the same underlying quantity, $\theta = 4N_e\mu$. In each population, which one is larger? Given that π weights by frequency while θ_W only counts sites, what does a gap between them tell you about whether the variants are common or rare?

For pop 1, $\pi > \theta_w$. It also has mean $H_o <$ mean H_e.
Pop 2 has $\theta_w > \pi$. It has mean $H_e >$ mean H_o too.

We expect rare variants to increase θ_w since it depends only on the count of sites while π is weighted by the number of comparisons. This might not be obvious from the limited example, but is provable.