DNA Sequence Variation

Alan R. Rogers

January 7, 2024

```
                0000000001 1111111112 2222222223 3333333334
                1234567890 1234567890 1234567890 1234567890

Sequence01      AATATGGCAC CTCCCAACCC TCTAGCATAT ACCACTTACA
Sequence02      .......T.. .C......TG C......C.. ..........
Sequence03      ..C....... .......... .......... ..........
Sequence04      .......T.. .C......TG C......... G.........
Sequence05      .......... .......... .......... ..........
Sequence06      .....A.... ........T. C......... G....C....
Sequence07      ..C....T.. .C......TG C......... G.........
Sequence08      .....A.T.. TC......TG C......... G.........
Sequence09      .......... .......... C......... ..........
Sequence10      .G...A.... ........T. C......C.. .T....C..G
Segregating:     ^^   ^ ^   ^^       ^^ ^       ^   ^^   ^^  ^
```

15 segregating sites

Measures of variation among DNA sequences

Gene diversity per sequence (a.k.a. heterozygosity) The probability that two random sequences differ.

Number, S, of segregating sites A "segregating site" is one that is polymorphic in the data.

Mean pairwise difference, Π, per sequence The average number nucleotide site differences between pairs of sequences.

Mean pairwise difference, π, per nucleotide Equals $\pi = \Pi/L$, where L is sequence length.

Mismatch distribution A histogram whose ith entry is the number of pairs of sequences that differ by i sites.

Site frequency spectrum A histogram whose ith entry is the number of polymorphic sites at which the mutant allele is present in i copies within the sample.

The number, $\binom{k}{2}$, of ways to choose 2 items from k

There are k ways to choose the first item. Having chosen the first, there are $k-1$ ways to choose the second, so there are $k(k-1)$ pairs. But this counts pair AB separately from BA. We are interested in unordered pairs, so

$$\binom{k}{2} = k(k-1)/2$$

A a set of made-up DNA sequences

```
          00000 00001
          12345 67890
S1        AAACT GTCAT
S2        ..... A....
S3        ..... A...C
S4        ..G.. A....
S5        ..G.. A....
```

Calculate the mean pairwise difference, the number of segregating sites, the mismatch distribution and the site frequency spectrum.

Mean pairwise difference (MPD)

```
    00000 00001        Pair Diff   Pair Diff
    12345 67890       (1,2)   1   (2,5)   1
S1  AAACT GTCAT       (1,3)   2   (3,4)   2
S2  ..... A....       (1,4)   2   (3,5)   2
S3  ..... A...C       (1,5)   2   (4,5)   0
S4  ..G.. A....       (2,3)   1   Sum diffs: 14
S5  ..G.. A....       (2,4)   1   MPD/seq  : 14/10
```

Column	Differences
03	2×3
06	1×4
10	1×4
Sum	14

Number of pairs: $(5 \times 4)/2 = 10$

MPD per sequence: $\Pi = 14/10$

MPD per site: $\pi = 14/(10 \times 10)$

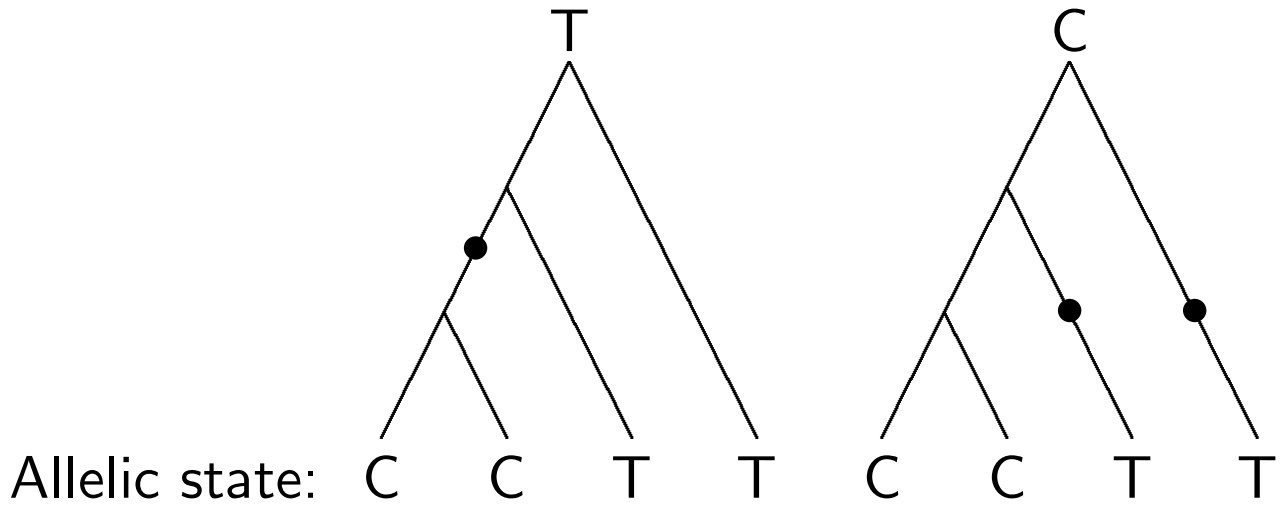
Mismatch distribution

```
    00000 00001        Pair Diff    Pair Diff
    12345 67890       (1,2)   1    (2,5)   1
S1  AAACT GTCAT       (1,3)   2    (3,4)   2
S2  ..... A....       (1,4)   2    (3,5)   2
S3  ..... A...C       (1,5)   2    (4,5)   0
S4  ..G.. A....       (2,3)   1
S5  ..G.. A....       (2,4)   1
```

Mismatch distribution

Differences	0	1	2
Count	1	4	5

Calling ancestral and derived alleles



- Two hypotheses about which allele is ancestral.
- "C" requires 2 mutations; "T" requires 1.
- Because mutations are rare, "T" is more likely.
- When the in-group is polymorphic, the ancestral allele is usually the one present in the out-group.

Unfolded site frequency spectrum

```
        123456
Human1  AATAGC
Human2  ..AC..
Human3  .TACT.
Human4  ..ACT.
------------
Chimp   AAAATC
```

1: fixed.
2: T derived; singleton.
3: T derived; singleton.
4: C derived; tripleton.
5: G derived; doubleton.
6: fixed.

Singletons	2
Doubletons	1
Tripletons	1