Appendix

**Frequency and distance between a detected allele and undetected alleles**

Frequency of detected alleles can be directly obtained from the result of KDE. When the component of the coordinate of an undetected allele is not “Inf”, then the frequency of the undetected allele is also directly obtained as well as the distance between the detected allele and undetected allele. The problem is when the corresponding component is recorded as “Inf”. In this case, both of the frequency and distance cannot be directly obtained and the following shows how to obtain them.

Consider a group of alleles which are with the length and distance away from an allele A and name the group as . The number of alleles belonging to is referred as , and Next, consider another allele B with the length that is distance away from the allele A, and name a group of alleles , which belong to and are distance away from the allele B (Figure S1). The number of alleles belonging to is referred as . Here, the allele B corresponds to a detected alleles and corresponds to undetected alleles.

To start from the simplest situation, consider . The allele A has bases, and of them are different bases from those of the allele B. Inevitably, of them are the same as the allele B. When a base in the allele A which is the same as the allele B mutates, the mutated allele is distance away from the allele B, and the total possible number of the mutated alleles is because there are three different bases except the one before mutation. When a base in the allele A which is different from that of the allele B mutates, there are two patterns: When the mutated allele is the same as the allele B, the mutated allele is distance away from the allele B, and its total possible number is . When the mutated allele is still different from that of the allele B, the mutated allele is distance away from the allele B, and its total possible number is . The above discussion is concisely expressed as the following equations:

Because distance between alleles changes one by one, , and exhaustively correspond to all the , which equals to .

Next, can be calculated from the result of . In , bases are the same as the allele B and bases are different from those of the allele B. When the former mutates, the mutated alleles are distance away from the allele B. When the later mutates, again there are two patterns: When the base after mutation is the same as the allele B, the mutated alleles are distance away from the allele B. When the base after mutation is different from that of the allele B, the mutated alleles remains distance away from the allele B. This is expressed as the following equations:

Likewise, total of the mutation in , and will be the following equations:

As shown above, the coefficient changes depending on the number of bases which are the same/different as/from that of the allele B. Now, mutation of a base in the allele A which is the same as the allele B is referred as , mutation of a base in the allele A which is different from that of the allele B into a base different from that of the allele B is referred as , and mutation of a base in the allele A which is different from that of the allele B into a base same as the allele B is referred as . Then, shown above can be expressed by the following way:

The multiplication among , and is not commutative, and they correspond to the following coefficients:

Here, is the number of before the corresponding multiplication, and is the number of the sum of and before the corresponding multiplication.

Now, total numbers of , and are referred as , and . Then, the following equations are satisfied:

Using the above expressions, sum of all the multiplications of , and provides. For example, the multiplication contributes to and the contribute is .

Since the coefficients , and are not commutative, the results of their multiplications are not single; the permutation of , and is

Therefore, will be expressed by the following equation:

Here, is when is even and is when odd.

This provides the number and distance from the allele B of undetected alleles around the allele A. Therefore, even when a coordinate of undetected alleles is “Inf” distance from the detected allele B, the frequency of the coordinate of undetected alleles can be distributed to each group of alleles having actual finite distance to the allele B proportional to the number of undetected alleles. Scanning from 1 to the mutation number and scanning all pairs of the detected allele A and B provide all the terms between a detected allele and undetected alleles to calculate approximate nucleotide diversity.