Additional file 1: Pseudocode for assessing variation

We start with a t samples comprising mapped data, a set of samples we call X.

Each sample in X is of length L. The objective is to maintain a vector variant_sites of length L containing either the invariant base present at each position (one of A,C,G,T) or a character (‘.’) indicating variation. On addition of the t^{th} sample $X_{t:}$

Pseudocode:

```
bases = [A,C,T,G]

alphabet = [N]

for i := 1 to L do

    if variant_sites[i] is not ‘.’ and $X_t$[i] in bases and variant_sites[i] != $X_t$[i] then

        variant_sites[i] = ‘.’

end
```