Supplementary Material

Given two STs that form an SLV, we assume a random variable, H, represents the number of base pairs that differ between the two STs. We are interested in the probability that an SLV at locus i with h differences is due to mutation only, that is, $Pr(h|SLV_i; M)$, where the event M=$\{$differences due only to mutation$\}$. M^c is the complement of M, i.e. the event of recombination(s) as well as mutations.

1 The model for differences due to mutation only

Firstly, the $Pr(h|SLV_i; M)$ was estimated according to Bayesian theory,

$$Pr(h|SLV_i; M) \propto Pr(h|M) \times Pr(SLV_i|h; M), \tag{1}$$

where $Pr(SLV_i|h; M)$ represents the likelihood function and $Pr(h|M)$ is the prior distribution of the number of nucleotide differences for the mutation only model.

Assume we have a mutation rate θ_i and recombination rate ρ_i at locus i. Denote $\theta = \sum_{i=1}^{7} \theta_i$ and $\rho = \sum_{i=1}^{7} \rho_i$. We set w_i=θ_i/θ, so w_i is the probability that if a mutation occurs it occurs at locus i. We estimate the w_is through estimating θ_is by the average number of base-pair differences between all alleles at locus i. Then we model the probability of an SLV at locus i given h base-pair differences, and that the SLV is caused only by mutation, as

$$Pr(SLV_i|h; M) \quad = \quad (w_i)^h. \tag{2}$$

This comes from the need that given M, there have been h mutations, and for it to be an SLV all mutations must occur at the same locus (locus i).

Finally, from coalescent theory we model that the probability of h mutations, given that there have only been mutations prior to the common ancestor of the pair of isolates, is geometric with parameter $\lambda = \theta/(1 + \rho + \theta)$. Thus we model

$$Pr(h|M) \quad \propto \quad Geometric(1 - \lambda), \tag{3}$$

and make the simplifying assumption that $\theta \approx \rho$ and $\theta, \rho >> 1$, thus, we have $\lambda \approx 0.5$. We use this value of λ in our analysis, but also considered how robust the results were to varying $\lambda < 0.5$ (as it appears that if anything $\rho > \theta$). Supplementary Table 1 and Supplementary Table 2 show that the choice of λ does not have a large effect on the results, with different choices of λ giving larger estimates for the relative rate of recombination to mutation.

We repeated our analysis under the assumption of equal mutation rate at each locus, and got very similar results for both *C. jejuni* and *C. coli*.

2 The recombination related model

The first step is to draw two alleles in that locus randomly based on the frequency of these alleles in one locus in PubMLST. The second step is to compare this pair of alleles and record the number of differences. This step was repeated for 1,000,000 iterations to obtain a stable

empirical probability distribution for observing h differences due to recombination for this locus: $Pr(h|M^c)$.

A naive approach to estimating the probability of observing h nucleotide differences being introduced by events that include recombination would be:

$$Pr(h|M^c) = \frac{n_h}{n_d} \tag{4}$$

in which h represents the number of nucleotide differences; n_h represents the item count of h differences (how many times h differences appears), n_d represents the number of all differences $n_d = \sum_{h=1}^{a} n_h$, where a is the maximum observed number of differences between any pair of alleles for the locus under consideration. However this is not robust, and the reason is that there are some values (say 30 to 45) of h for which $n_h = 0$, i.e. pairs of alleles with 30 to 45 differences are never observed in the sample. Using Equation 4 would then estimate the probability of recombination producing such a number of differences as 0; and if we observe h differences in our SLVs data, our model would have to assign this to mutations. A simple way around this is to introduce a Dirichlet prior on $Pr(h|M^c)$, which gives the posterior estimates:

$$Pr(h|M^c) \;=\; \frac{1 + n_h}{a + n_d}. \tag{5}$$

3 Mixture model

For this part, we will consider a fixed locus i; and estimate the probability that an SLV was the result of mutation only for that locus (p_i). It was assumed that there are data h_1, h_2, h_3,..., $h_{n_{data}}$, where n_{data} is the number of pairs of distinct STs with SLVs in PubMLST dataset, and h_j ($j = 1, 2, ..., n_{data}$) represents the number of nucleotide differences of the jth SLV.

The distribution for h for an SLV is

$$f(H) = \begin{cases} Pr(H|SLV_i, M), & if\ z = 1 \\ Pr(H|SLV_i, M^c), & if\ z = 0 \end{cases} \tag{6}$$

in which $Pr(H|SLV_i, M)$ is the model for solely mutation, and $Pr(H|SLV_i, M^c)$ is the recombination related model. The latent variable Z is introduced as the indicator to tell whether the data h_j comes from either of these two models. For example, when $z_j = 1$, h_j comes from the mutation model; whereas when $z_j = 0$, h_j comes from the recombination model. Thus the probability of $z_j = 1$ is the proportion of SLVs caused solely by mutation (p_i), and

$$f(h_j,\, z_j|SLV_i) = p_i \times Pr(h_j|SLV_i, M) + (1 - p_i) \times Pr(h_j|SLV_i, M^c). \tag{7}$$

We can estimate p_i under this model by maximum likelihood using the EM Algorithm. The expectation-maximization (EM) algorithm is an approach for finding maximum-likelihood estimates by iterative computation, when the statistical model depends on unobserved latent variable (Dempster et al., 1977). It includes two steps: expectation step (E-step) and maximization step (M-step). In the E-step, the expectation of log likelihood was calculated, in the M-step, the expectation was maximized. 100 iterations were run to get stable parameter estimates, although the parameters converged after 50 iterations.

From the EM algorithm, we get:

$$\hat{p}_{i,new} = \frac{\mathrm{E}(M|p_{i,old})}{n}, \tag{8}$$

in which

$$\mathrm{E}(M|p_{i,old}) = \sum_{j=1}^{n} Pr(z_j = 1|p_{i,old}). \tag{9}$$

4 The model for an event being mutation rather than recombination

SLVs can be caused by multiple events. Let K be the number of events separating the two branches in the evolutionary tree of each locus. Coalescent theory gives that the number of events (mutation/recombination) between two randomly chosen isolates follows a geometric distribution with parameter $1/(1+\rho+\theta)$:

$$Pr(k) = \left(\frac{\rho+\theta}{1+\rho+\theta}\right)^k \left(\frac{1}{1+\rho+\theta}\right) \tag{10}$$

The probability of an SLV at locus i given $K = k$ is

$$Pr(SLV_i|k) = \left(\frac{\rho_i+\theta_i}{\rho+\theta}\right)^k.$$

Thus,

$$Pr(k|SLV_i) \propto Pr(SLV_i|k) \times Pr(k) \tag{11}$$

$$\propto \left(\frac{\rho_i+\theta_i}{\rho+\theta}\right)^k \left(\frac{\rho+\theta}{1+\rho+\theta}\right)^k \tag{12}$$

$$\propto \left(\frac{\rho_i+\theta_i}{1+\rho+\theta}\right)^k, \tag{13}$$

in which, $Pr(SLV_i|k)$ means the probability that SLVs at locus i are caused by k events. Assuming $\theta+\rho >> 1$ and ρ_i is roughly proportional to θ_i we then have the approximation

$$Pr(K = k|SLV_i) \propto \omega_i^k. \tag{14}$$

As $Pr(k|SLV_i)$ is a probability mass function we obtain

$$Pr(k|SLV_i) = (1-\omega_i) \times (\omega_i)^{k-1}, \text{for } i = 1, 2, \ldots. \tag{15}$$

Now, p_i has been defined as the probability of an SLV involves only mutation and no recombination at locus i, while x_i is defined as the actual probability that an event for generating new alleles that led to SLVs is mutation rather than recombination at locus i. The relationship between p_i and x_i is thus

$$p_i = \sum_{k=1}^{\infty} (1-\omega_i) \times (\omega_i)^{k-1} x_i^k, \tag{16}$$

where the sum is over the number of events, and we need all events to be mutations. Thus we get that $x_i = p_i/(1-\omega_i + p_i \times \omega_i)$. Then $(1-x_i)/x_i$ represents the relative rate of recombination to mutation.

Using the same model we obtain an estimate of the number of mutation events at an SLV at locus i. From Equation 15 we have the expected number of events is $\mathrm{E}(K|SLV_i) = 1/(1-\omega_i)$, and a proportion x_i of all such events are mutations. Thus the expected number of mutation events is $x_i/(1-\omega_i)$. The proportion of nucleotide differences of an SLV due to recombination is calculated by $1 - x_i/(d - d \times \omega_i)$, d is the average number of differences in all SLVs at locus i.

5 The comparison with Feil et al.'s method

A simplified version of Feil et al.'s (2000) method that assumed that all differences of one nucleotide were caused by mutation, but larger nucleotide differences were due to recombination was also applied.

Results (Supplementary Figure 1) show, for the simplified version of Feil et al.'s method, the ratio of recombination to mutation were overestimated, compared to our results. As Feil et al.'s full method has the potential to underestimate mutation even more, their estimation of the ratio of recombination to mutation will be apparently higher than our estimates.

References

Dempster, A., Laird, N., Rubin, D., et al. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Feil, E., Smith, J., Enright, M., and Spratt, B. (2000). Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. *Genetics*, 154(4):1439.