The Annals of Applied Statistics
2015, Vol. 9, No. 1, 200–224
DOI: 10.1214/14-AOAS795

ESTIMATING THE RELATIVE RATE OF RECOMBINATION TO MUTATION IN BACTERIA FROM SINGLE-LOCUS VARIANTS USING COMPOSITE LIKELIHOOD METHODS[1]

BY PAUL FEARNHEAD[2,*], SHOUKAI YU[†], PATRICK BIGGS[†], BARBARA HOLLAND[‡] AND NIGEL FRENCH[†]

Lancaster University[*], *Massey University*[†] *and University of Tasmania*[‡]

A number of studies have suggested using comparisons between DNA sequences of closely related bacterial isolates to estimate the relative rate of recombination to mutation for that bacterial species. We consider such an approach which uses single-locus variants: pairs of isolates whose DNA differ at a single gene locus. One way of deriving point estimates for the relative rate of recombination to mutation from such data is to use composite likelihood methods. We extend recent work in this area so as to be able to construct confidence intervals for our estimates, without needing to resort to computationally-intensive bootstrap procedures, and to develop a test for whether the relative rate varies across loci. Both our test and method for constructing confidence intervals are obtained by modeling the dependence structure in the data, and then applying asymptotic theory regarding the distribution of estimators obtained using a composite likelihood. We applied these methods to multi-locus sequence typing (MLST) data from eight bacteria, finding strong evidence for considerable rate variation in three of these: *Bacillus cereus*, *Enterococcus faecium* and *Klebsiella pneumoniae*.

1. Introduction. Homologous recombination is a process which allows foreign DNA to be incorporated within a genome. In bacteria this can occur through three different mechanisms: conjugation, the uptake of DNA from other bacteria; transformation, the uptake of naked DNA from the remains of bacteria that exist in the living environment; or transduction, where DNA is implanted by bacteriophage [Low and Porter (1978)]. Although different, each result in the Introduction of a new DNA sequence within a region of the genome, and thus recombination is potentially an important mechanism driving the evolution of a given bacteria. Understanding recombination in bacteria is important because it can allow for genetic exchange between distant bacterial species and impacts on the evolution of

Received September 2013; revised June 2014.
[1]Supported in part by the Marsden Fund Project 08-MAU-099 (Cows, starlings and Campylobacter in New Zealand: unifying phylogeny, genealogy and epidemiology to gain insight into pathogen evolution).
[2]Supported in part by the Engineering and Physical Sciences Research Council, UK, Grant EP/K014463/1.
Key words and phrases. Composite likelihood, recombination, single-locus variants, testing for rate variation.

new species [Fraser, Hanage and Spratt (2007), Sheppard et al. (2008)]. Furthermore, the rate of recombination varies considerably across bacterial species: with estimates of the relative effect of recombination to mutation varying by over three orders of magnitude in Vos and Didelot (2009). Here we look at how to estimate the relative rate of recombination to point mutation from population genetic data that describe the genetic variation between a sample of bacterial isolates at a number of loci. In particular, our approach develops recent ideas that estimate this relative rate by comparing the DNA for closely related isolates.

For population genetic data it is often helpful to consider the genealogical history of a sample. If there is no recombination, this can be represented by a single tree, which is often called the genealogy of the sample. The effect of recombination is that, while at any specific position along the chromosome we can define such a genealogical tree, this tree can be different for different positions. The genealogical history of a sample is thus defined by the collection of all such trees, which can be represented through a graph [Griffiths and Marjoram (1997)].

Within bacteria each recombination event generally affects only a relatively small region of the genome. For example, in *Campylobacter jejuni*, a recombination event may change the DNA within a region of between a few hundred to a few thousand base pairs, which constitutes a fraction of a percent of the whole 1.6 Mb genome. Thus, we can define a single tree for a sample of bacteria by tracing the ancestry of the nonrecombinant region at each recombination event. This tree has been called the clonal frame [Didelot and Falush (2007), Milkman and Bridges (1990)]. We can then model recombination events as introducing a number of mutations onto this clonal frame. An example is given in Figure 1. Our approach to estimating recombination rates in this paper is based on using such a model.

In this paper we assume we have genetic data collected from a number of isolates of a given bacteria, and that this genetic data consist of the DNA sequence at L loci of similar size. We assume these loci are sufficiently spread around the genome such that a single recombination event is unlikely to affect more than one locus. An example of data satisfying these assumptions is MLST data [Maiden et al. (1998)], which consist of the DNA sequence of $\approx$500 bp fragments from a selection of, normally around 7, housekeeping genes. Large MLST data sets for over 20 bacteria are available from http://pubmlst.org.

For such data we can define *sequence types (STs)* so that two isolates which have identical sequences at all L loci will have the same ST, but any two isolates whose sequences differ will have differents STs. It is standard to define STs numerically: ST1, ST2 etc. A simple example for 3 loci and 7 bacterial isolates is given in Figure 2, where we also show the underlying clonal frame of the sample, and the mutation and recombination events that have affected the sample. If we assume each mutation is distinct, and these are also different from the mutations introduced at recombination events, then we get 6 distinct sequence types.

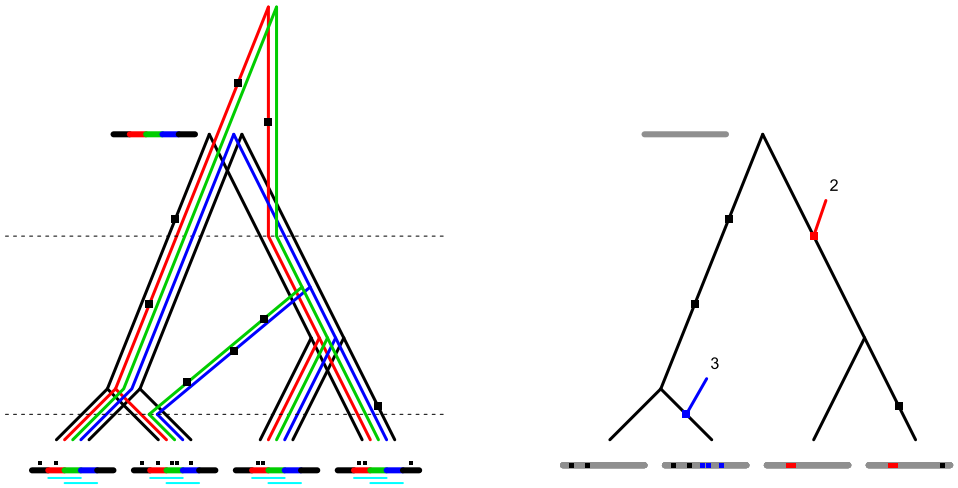

FIG. 1. *Left-hand plot*: *Example of the genealogical trees for* 4 *isolates in one region of the genome. The effect of recombination is that different sites in the gene have different genealogical trees. Here we have two recombination events that have affected the genealogical relationship of the isolates* (*the position of these back in time are denoted by the dashed lines*, *and the regions they affect by the light blue lines under the gene fragments*). *The black tree is the clonal frame of the sample*: *the genealogy at regions unaffected by recombination. Mutations that have affected the sample are given by the black squares. Here we consider mutations that create differences from the sequence of the common ancestor on the clonal frame. The other trees represent the genealogies of regions affected by either one or both recombination events. Right-hand plot*: *the simplified model for the data based on the clonal frame. Differences within our sample are created by mutation and recombination events that occur on the clonal frame. We do not track the ancestry of recombination events*, *instead each one just introduces a number of mutations within the recombinant region. These events are shown in the figure and are labeled with the number of differences introduced.*

The single-locus variants (SLVs) of a specific ST will be the set of other STs that differ from it only at a single-locus. If we consider pairs of SLVs at a specific locus l, then an SLV pair will be defined as a pair of isolates that have different DNA sequences at locus l but have identical DNA sequences at the other $L - 1$ loci. For the example in Figure 2 we get one SLV pair at locus 2, and 3 SLV pairs at locus 3. These can be summarized by the STs of each pair together with the number of nucleotide differences at the locus that differs; see Table 1. In this paper we consider how to use data such as that in Table 1 to infer the relative rate of recombination to mutation. Note that we define this relative rate as the ratio of the rate at which a locus is affected by recombination to the rate at which it is affected by mutation. This rate is different from the relative rate of recombination to mutation events across the genome, as recombination events that start outside a locus can still affect it. Thus, the rate at which a locus is affected by recombination depends both on the rate of recombination events and the relative size of the average recombination tract length to the size of the locus.

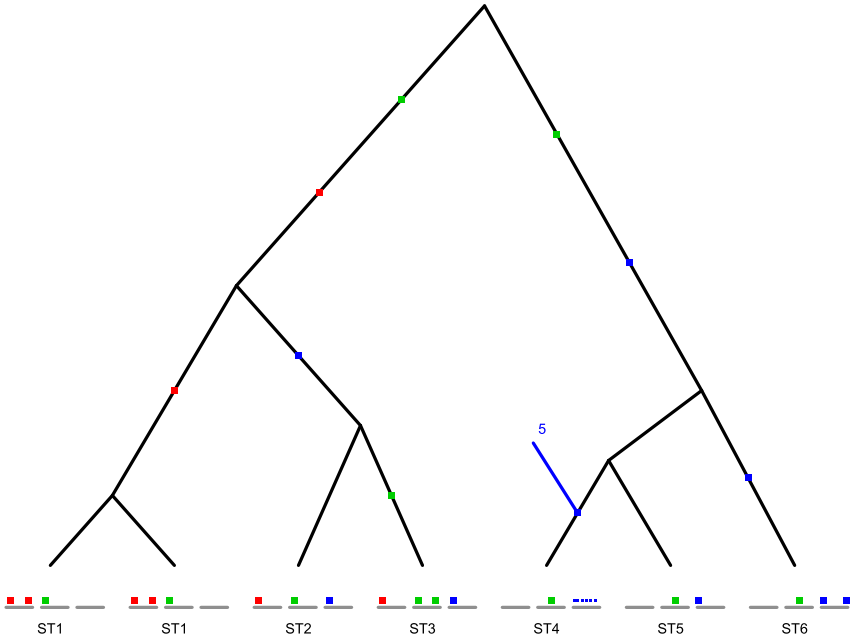

FIG. 2. *Example clonal frame and resulting data set for* 7 *isolates at* 3 *loci. Mutations are denoted by squares, and the color denotes the locus that the mutation occurs on. There is a single recombination event, denoted by a square labeled with the number of differences to the ancestral sequence that event introduces. For this example, there are* 6 *distinct sequence types, denoted ST*1 *to ST*6*. We have an SLV pair at locus* 2 (*ST*2 *and ST*3) *and* 3 *SLV pairs at locus* 3 (*all distinct pairs of ST*4 *to ST*6).

The idea of using SLVs to estimate the relative rate of recombination to mutation comes initially from the work of Feil et al. (1999) [see also Feil et al. (2000), Spratt, Hanage and Feil (2001)]. By comparing closely related isolates, it can often be clear as to whether the pair of isolates differs only by mutation or not. The approach in Feil et al. (1999) is based on assuming that if the number of nucleotide

TABLE 1

Data for SLV pairs from example in Figure 2*. The data consist of the SLV pairs for each locus, together with* x*, the number of nucleotide differences each SLV pair has at that locus*

Locus	**ST**	**ST**	x
2	2	3	1
3	4	5	5
3	4	6	6
3	5	6	1

differences is small (say 2 or fewer), then these are caused by mutation. If the number of differences is large, then these are caused by recombination. For the data in Table 1 such an approach would work well: identifying correctly that two of the four SLV pairs are created by mutation, and two involve recombination.

However, to obtain sensible estimates of the relative rate of recombination to mutation, we need to deal with two issues. First is the fact that SLV pairs that involve recombination may have differences caused by both mutation and recombination. Thus, using a simple ratio of SLV pairs caused by only mutation to those that involve recombination may not be appropriate. Second, some recombination events may introduce a small number of nucleotide differences, and thus we need to allow for some of the SLV pairs that differ at a small number of nucleotides to be due to recombination. To address these issues, Yu et al. (2012) introduce a simple model for the number of nuceotide differences for an SLV pair as a function of the relative rate of recombination to mutation, and then estimate the parameter in this model using composite likelihood. We take the same approach here.

While the model is approximate, it should give robust and accurate inferences in situations where it is easy to detect whether SLVs are caused only by mutation and where it is likely that SLVs are caused by only one, mutation or recombination, event. This corresponds to SLVs defined for data collected at a relatively large number of loci (such as the 7 used in MLST data), and where most recombination events introduce a large number of nucleotide differences.

A second issue with using SLVs to estimate the relative rate of recombination to mutation is that data from some SLV pairs are dependent. This can be seen in the data in Table 1 and the three SLV pairs at locus 3. These three SLV pairs are caused by two events: one mutation and one recombination. This dependence makes it harder to assess uncertainty in parameter estimates. As a result, Yu et al. (2012) resort to using simulation, via a parametric bootstrap, to assess this uncertainty. The main disadvantage with this is that the accuracy of the resulting measures of uncertainty will depend on how accurate the model used to simulate the data is and, in practice, the models used to simulate data do not capture many of the features observed in real data sets. Furthermore, the use of simulation adds substantially to the overall computational effort required to analyze any data set: using a parametric bootstrap for large data sets, such as the *Staphylococcus aureus* MLST data set we analyze in Section 4, can take months of CPU time.

The main difference between this work and that of Yu et al. (2012) is that we use theory for composite likelihoods to get direct measures of uncertainty of our estimates. We believe these to be more reliable than using simulation, and they are much quicker to calculate. Composite likelihood theory also gives a framework for performing inference across loci. We show how we can test whether there are differences in the relative rate of recombination to mutation across loci, and how to get an estimate of the common relative rate under the assumption that there is no variation across loci.

The next section introduces our model for data from SLV pairs and how we can use composite likelihood to estimate the relative rate of recombination to mutation from such data and to quantify the uncertainty in these estimates. We evaluate this method on simulated data in Section 3. Our results suggest that we can get both accurate estimates and also appropriate measures of uncertainty of our estimators. For large data sets, coverage values of our confidence intervals do drop away from their nominal value, but this appears to be due to slight biases in our model as opposed to underestimating the variability of the estimators. We show that tests for detecting differences across loci have close to their nominal significance level across a range of simulated scenarios, and have good power to detect rate variation across loci in situations where the recombination varies by a factor of three or more. In Section 4 we apply our method to analyze MLST data from 8 bacteria. In three cases we find strong evidence that the rate of recombination to mutation varies across loci, with estimates suggesting this variation could be by up to two orders of magnitude. The paper ends with a discussion. Data and code used in the paper are available at http://www.maths.lancs.ac.uk/~fearnhea/SLV.zip.

2. Estimating recombination rates from SLVs. We will now describe how we use data like that in Table 1 to estimate the relative rate of recombination and mutation at a locus using composite likelihood methods. Roughly, the idea behind composite likelihood approaches is to (i) split data into small subsets; (ii) introduce a probabilistic model for each subset of data, which in turn will define a log-likelihood for that subset; and (iii) combine information from all subsets through taking a weighted sum of the log-likelihoods from the subsets. The weighted sum is called a composite likelihood. Parameters can be estimated through maximizing this composite likelihood. Furthermore, if we model the dependence between the log-likelihoods across different subsets, we are able to estimate the asymptotic variance of, and construct confidence intervals for, the resulting parameter estimates.

Implementing a composite likelihood method involves a number of key decisions and modeling assumptions. The first is the choice of subsets of the data to use. In our application each subset corresponds to an SLV pair, with the data being the number of nucleotide differences we observe for that SLV pair. Second, we need to develop a model for the data, and we describe our model in Section 2.1. Then we need to choose the weights used when constructing the composite likelihood, and finally to choose how to model the dependence between subsets of the data. The last aspect is needed in order to be able to assess the uncertainty of estimators, and hence to obtain confidence intervals. One advantage of composite likelihood methods is that we need only model one aspect of this dependence, namely, the covariance of the score function (the derivative of the log-likelihood) for pairs of subsets of data. Our choices of weights and model for this dependence for our application are introduced in Section 2.2, together with fuller details of

how we then construct confidence intervals for our estimate of the relative rate of recombination and mutation.

Theory for composite likelihood can also be used to combine information across loci and construct a test for whether the relative rate of recombination to mutation varies across loci. This is described in Section 2.3.

2.1. *Model for a single SLV pair.* Here we derive a conditional likelihood for data from a single SLV pair at a specific locus, l, say. The data are the number of nucleotide differences that the SLV pair has at that locus, which will be denoted x. Remember that the total number of loci is L, and we will denote the relative rate at which recombination, as opposed to mutation, affects locus l by λ. Our conditional log-likelihood will be based on the log of the probability of observing x differences between two STs conditonal on these STs being an SLV pair at locus l,

$$\ell(\lambda; x) = \log \Pr_{\lambda}(X = x | \text{SLV at locus } l),$$

where the random variable X denotes the number of nucleotide differences at locus l between a (random) pair of isolates, and we are conditioning on the pair of isolates being an SLV pair at locus l. We include the subscript λ as the probability depends on λ.

To calculate this conditional probability, let A denote the event that only mutations have occurred at locus l to create the differences between the pair of isolates, and A^c the complementary event that at least one recombination occurred at locus l. Further, let θ_l denote the mutation rate at locus l, and $\theta = \sum_{i=1}^{L} \theta_i$ denote the overall mutation rate across the L loci. [We use standard coalescent scaling for these rates, so one unit of time is equal to the expected time in the past until a pair of isolates share a common ancestor; see Wakeley (2007).] Finally, we shall assume that the relative rate of recombination to mutation, λ, is the same across all loci.

Under a standard coalescent model, the probability of a pair of isolates being an SLV pair at locus l is

$$\begin{aligned}
\Pr_{\lambda}(\text{SLV at locus } l) &\approx \frac{1}{1 + (1+\lambda)\theta} \sum_{i=1}^{\infty} \left(\frac{(1+\lambda)\theta_l}{1 + (1+\lambda)\theta} \right)^i \\
&\approx \frac{1}{(1+\lambda)\theta} \sum_{i=1}^{\infty} \left(\frac{(1+\lambda)\theta_l}{(1+\lambda)\theta} \right)^i \\
&= \frac{1}{\theta(1+\lambda)} \frac{\theta_l}{\theta} \left(1 - \frac{\theta_l}{\theta} \right)^{-1} = \frac{1}{\theta(1+\lambda)} \frac{\theta_l}{\theta - \theta_l}.
\end{aligned}$$

The first approximation comes from assuming that all recombination events at a locus introduce a change to the sequence of one of the isolates. The expression on the right-hand side is then obtained by considering the possible events in the

history of the two isolates back to their common ancestor: there will need to be at least one mutation/recombination event at locus l, and no mutation/recombination events at other loci. The $1/(1+(1+\lambda)\theta)$ is the probability that the next event back in time is a coalescent event, and the $(1+\lambda)\theta_l/(1+(1+\lambda)\theta)$ is the probability of the next event back in time being a mutation or recombination event at locus l [Wakeley (2007)]. We sum over i, the number of mutation or recombination events in the history of the SLV pair. The approximation we have used in the second line is reasonable if $(1+\lambda)\theta \gg 1$, which will be true for the cases where this approach to inference for λ can be expected to be accurate. The advantage of this approximation is that it means our final expression for the conditional likelihood (see below) will only depend on θ_l and θ through the ratio θ_l/θ, which will be easier to estimate in practice.

Now, by a similar argument, and using the same approximation, we can get

$$\Pr_{\lambda}(X = x \cap A \cap \text{SLV at locus } l) \approx \frac{1}{\theta(1+\lambda)}\left(\frac{\theta_l}{\theta(1+\lambda)}\right)^x,$$

as this will require x mutation events at locus l followed by a coalescent event. Summing over $x = 1, 2, \ldots$ gives

$$\Pr_{\lambda}(A \cap \text{SLV at locus } l) \approx \frac{1}{\theta(1+\lambda)}\frac{\theta_l}{\theta+\lambda\theta-\theta_l}.$$

Thus, we can write

$$\begin{aligned}
&\Pr_{\lambda}(X = x|\text{SLV at locus } l)\\
&\quad\propto \Pr_{\lambda}(X = x \cap \text{SLV at locus } l)\\
&\quad= \Pr_{\lambda}(X = x \cap A \cap \text{SLV at locus } l) + \Pr_{\lambda}\big(X = x \cap A^c \cap \text{SLV at locus } l\big)\\
&\quad\approx \frac{1}{\theta(1+\lambda)}\left(\frac{\theta_l}{\theta(1+\lambda)}\right)^x\\
&\qquad+ \Pr_{\lambda}\big(\text{SLV at locus } l \cap A^c\big)\Pr_{\lambda}\big(X = x|\text{SLV at locus } l, A^c\big)\\
&\quad= \frac{1}{\theta(1+\lambda)}\left(\frac{\theta_l}{\theta(1+\lambda)}\right)^x\\
&\qquad+ \Big(\Pr_{\lambda}(\text{SLV at locus } l) - \Pr_{\lambda}(\text{SLV at locus } l \cap A)\Big)\\
&\qquad\times \Pr_{\lambda}\big(X = x|\text{SLV at locus } l, A^c\big)\\
&\quad\approx \frac{1}{\theta(1+\lambda)}\left(\frac{\theta_l}{\theta(1+\lambda)}\right)^x\\
&\qquad+ \left(\frac{1}{\theta(1+\lambda)}\frac{\theta_l}{\theta-\theta_l} - \frac{1}{\theta(1+\lambda)}\frac{\theta_l}{\theta+\lambda\theta-\theta_l}\right)
\end{aligned}$$

$$\times \Pr_{\lambda}(X = x|\text{SLV at locus } l, A^c)$$

$$\propto \left(\frac{\theta_l}{\theta(1+\lambda)}\right)^x + \left(\frac{\theta_l}{\theta - \theta_l} - \frac{\theta_l}{\theta + \lambda\theta - \theta_l}\right) \Pr_{\lambda}(X = x|\text{SLV at locus } l, A^c).$$

We can calculate the normalizing constant of this distribution by summing the final terms over all possible values for x.

To use this conditional likelihood, we need to specify the probabilities $\Pr_\lambda(X = x|\text{SLV}, A^c)$ and θ_l/θ. We approximate the former by the distribution of the number of nucleotide differences that are introduced by a single recombination event, on the basis that we expect the number of mutation/recombination events in the history of an SLV pair to be small, and the recombination event will produce most of the differences between the two isolates. This can then be empirically approximated based on simulating recombination events at locus l from data on the population diversity of sequences at that locus (see Appendix A). To estimate θ_l/θ, we can either use the relative size of the sequences at each locus (approximating the mutation rate per bp as constant) or we can use estimates of the mutation rate at each locus from population data, such as based on the mean number of pairwise differences [Donnelly and Tavaré (1995)]. Our experience is that the results are robust to the method used, and we suggest the former unless there is strong evidence that mutation rates vary substantially across loci.

This conditional likelihood is very similar to that derived in Yu et al. (2012). The main difference is that a further approximation is used in Yu et al. (2012) whereby the probability of two isolates being an SLV pair does not depend on λ.

2.2. *Composite likelihood inference.* We now consider how to estimate λ given data from a set of n SLV pairs. Denote the number of differences at each of the n SLV pairs by $\mathbf{x} = (x_1, \ldots, x_n)$. If the data from each SLV pair were independent from the others, then it would be natural to estimate λ by maximizing $\sum_{i=1}^n \ell(\lambda; x_i)$; however, the data from each SLV pair is not necessarily independent. To see this, consider the SLV pairs in Table 1. We have three SLV pairs involving each pair of ST4, ST5 and ST6. The data for these three SLV pairs are dependent, as different SLV pairs are affected by the same events. For example, both SLV pairs involving ST4 are affected by the same recombination event.

Despite the data being dependent, we can still estimate λ by maximizing a related function

$$\text{Cl}(\lambda; \mathbf{x}) = \sum_{i=1}^{n} w_i \ell(\lambda; x_i),$$

where $w_1, \ldots, w_n$ are a set of positive weights. In this case $\text{Cl}(\lambda; \mathbf{x})$ is often called a composite likelihood, and asymptotic theory exists which shows that in many situations maximizing a composite likelihood produces an estimator with good statistical properties, such as consistency and asymptotic normality [Varin, Reid and

Firth (2011)]. The choice of weights affects the overall accuracy of the estimator, and we suggest an appropriate choice for our application below. Furthermore, this theory shows how to construct appropriate confidence intervals for parameters. Here we outline one such result that we will use.

Assume the true parameter value is λ_0 and that $\hat{\lambda}$ maximizes $\mathrm{Cl}(\lambda; \mathbf{x})$. Define the score function

$$u(\lambda; x) = \frac{\mathrm{d}\ell(\lambda; x)}{\mathrm{d}\lambda} \quad \text{and}$$

$$U(\lambda; \mathbf{x}) = \frac{\mathrm{d}\,\mathrm{Cl}(\lambda; \mathbf{x})}{\mathrm{d}\lambda} = \sum_{i=1}^{n} w_i u(\lambda; x_i).$$

Define $J(\lambda) = \mathrm{Var}(U(\lambda; \mathbf{X}))$ and

$$I(\lambda) = -\mathrm{E}\left(\frac{\mathrm{d}U(\lambda; \mathbf{X})}{\mathrm{d}\lambda}\right),$$

where in each case we calculate the variance or expectation with respect to data sets $\mathbf{X}$ being drawn from the model with parameter λ. Then if we set $\gamma = J(\lambda_0)/I(\lambda_0)$, we can calculate a scaled deviance

$$W(\lambda) = \frac{2}{\gamma}\big[\mathrm{Cl}(\hat{\lambda}; \mathbf{x}) - \mathrm{Cl}(\lambda; \mathbf{x})\big].$$

Under certain regularity conditions, as $n \to \infty$, $W(\lambda_0)$ is asymptotically chi-squared distributed with 1 degree of freedom. If we can calculate, or consistently estimate γ, then this result can be used to construct a confidence interval for λ. Note that if $\mathrm{Cl}(\lambda; \mathbf{X})$ is replaced by a true log-likelihood, then, under standard regularity conditions, both $I(\lambda)$ and $J(\lambda)$ are equal to the Fisher information. In this case $\gamma = 1$.

Each $u(\lambda; X_i)$ is identically distributed. Let $\sigma^2 = \mathrm{Var}(u(\lambda_0; X))$, and note that standard results for the expected information give that $I(\lambda_0) = \sum_{i=1}^{n} w_i \sigma^2$. Now we can write

$$J(\lambda_0) = \sigma^2\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \,\mathrm{Cor}\big[u(\lambda; X_i); u(\lambda; X_j)\big]\right). \tag{1}$$

For any data set we will be able to partition the STs involved in SLV pairs at locus l into groups, so that if you take any pair of STs within a group, they will form an SLV pair, but if you take any two STs which come from different groups, they will not form an SLV pair. Assume that there are G groups, containing $n_1, \ldots, n_G$ STs, respectively.

A consequence of this is that you can also split the set of SLV pairs into the same number of groups: so the gth group of SLV pairs will consist of all $n_g(n_g - 1)/2$ pairs from the gth group of STs. The total number of SLV pairs will be $n = \sum_{g=1}^{G} n_g(n_g - 1)/2$. The dependence in the data, the number of nucleotide

differences of each SLV pair, is due to the possibility of different SLV pairs being affected by the same mutation and/or recombination events, as was seen in the example in Table 1. By construction, SLV pairs taken from different groups will not share any mutation or recombination events, hence, it is reasonable to assume the data from SLV pairs in different groups will be independent. For all distinct SLV pairs within the same group, the correlation between them will be the same (by symmetry). To develop a parsimonious model for the correlation structure within a group, we will make a number of further simplifying assumptions. First is that the correlation is the same for distinct SLV pairs as it is for SLV pairs that share a common ST. Second is that the level of correlation does not depend on the size of the group. Under these assumptions we will have for some $\alpha \in [0, 1]$,

$$\begin{aligned}&\operatorname{Cor}\big[u(\lambda; X_i); u(\lambda; X_j)\big]\\ &\quad= \begin{cases} 1, & \text{if } i = j,\\ \alpha, & \text{if } i \neq j \text{ but SLV pairs } i \text{ and } j \text{ are in the same group},\\ 0, & \text{otherwise.}\end{cases}\end{aligned}$$

We choose the weights based on this dependence structure. A group of SLV pairs based on n_g STs will contribute $n_g(n_g - 1)/2$ terms to the composite likelihood, but these depend on just n_g pieces of information (the data at the n_g different STs). If we chose uniform weights, then this would mean that the composite likelihood could be overly dominated by the data from a small number of large groups, which would lead to an increase of the variance of our estimate of λ. Thus, we choose to downweight SLV pairs that are in large groups. Our particular choice is that for an SLV pair in a group of size n_g we have $w_i = f_w(n_g) = \{n_g(n_g - 1)/2\}^{-1/2}$, the inverse of the square root of the number of SLV pairs in that group.

Substituting these definitions for the correlation and the weights into (1) gives

$$\begin{aligned}J(\lambda_0) &= \sigma^2\Bigg(\sum_{g=1}^{G} \frac{n_g(n_g-1)}{2} f_w(n_g)^2\\ &\qquad + \alpha \sum_{g=1}^{G} \frac{n_g(n_g-1)}{2}\left\{\frac{n_g(n_g-1)}{2} - 1\right\} f_w(n_g)^2\Bigg)\\ &= \sigma^2\Bigg(\sum_{g=1}^{G} 1 + \alpha \sum_{g=1}^{G}\left\{\frac{n_g(n_g-1)}{2} - 1\right\}\Bigg).\end{aligned}$$

Thus,

$$\frac{J(\lambda_0)}{I(\lambda_0)} = \frac{(G + \alpha \sum_{g=1}^{G}\{n_g(n_g-1)/2 - 1\})}{\sum_{g=1}^{G}\{n_g(n_g-1)/2\}^{1/2}}.$$

Note that this ratio does not depend on λ_0, it purely depends on the correlation parameter α. We can estimate α from the empirical correlation of the $u(\hat{\lambda}; x_i)$'s for SLV pairs within the same group. We do this formally by modeling the score functions as multivariate Gaussian with the appropriate covariance structure. In practice, we make the further assumption of a common α value for all loci. Details are given in Appendix B.

2.3. *Joint inference across loci.* It is possible to combine inferences across loci. It is natural to assume that, conditional on parameters, the data from one locus is independent of data from another, as the SLVs will be caused by different recombination and mutation events. Two natural questions to address from looking across loci are whether the value of λ is the same for all loci and, if so, can we estimate this common λ value. We will show how composite likelihood methods can be used to answer both these questions.

First we introduce some notation. Let $\mathrm{Cl}^{(l)}(\lambda)$ be the composite log-likelihood function for locus l. Similarly, let $J(\lambda)^{(l)}$ and $I(\lambda)^{(l)}$ denote the corresponding values of $J(\lambda)$ and $I(\lambda)$ for locus l.

We first consider answering the second question, and let λ_0 be the true common λ value for the loci. We can estimate this by maximizing the sum of the locus-specific composite log-likelihoods:

$$\hat{\lambda} = \arg\max \sum_{l=1}^{L} \mathrm{Cl}^{(l)}(\lambda).$$

Furthermore, if we define a scaled deviance as

$$W(\lambda) = \frac{2}{\gamma}\left[\sum_{l=1}^{L} \mathrm{Cl}^{(l)}(\hat{\lambda}) - \sum_{l=1}^{L} \mathrm{Cl}^{(l)}(\lambda)\right],$$

where $\gamma = \sum_{l=1}^{L} J^{(l)}(\lambda_0) / \sum_{l=1}^{L} I^{(l)}(\lambda_0)$, then $W(\lambda_0)$ asymptotically has a chi-squared distribution with one degree of freedom [Varin, Reid and Firth (2011)]. We can estimate γ using our estimates of $J(\lambda_0)^{(l)}$ and $I(\lambda_0)^{(l)}$ from each locus.

To test whether the value of λ at each locus is the same, we can use a (composite) likelihood-ratio test statistic. We define the test statistic to be proportional to the difference in composite log-likelihood for a model which allows each locus to have different λ values and that of a model with a common λ value across loci:

$$LR = \frac{2}{\nu_1}\left[\sum_{l=1}^{L} \max_{\lambda} \mathrm{Cl}^{(l)}(\lambda) - \max_{\lambda} \sum_{l=1}^{L} \mathrm{Cl}^{(l)}(\lambda)\right].$$

If our null hypothesis, of a common λ across loci, is true, then LR has an asymptotic distribution that is an inhomogeneous sum of independent chi-squared distributions [Varin, Reid and Firth (2011)]. For an appropriate choice of ν_1 it is common to approximate this asymptotic distribution for a chi-squared distribution

with $L-1$ degrees of freedom [Molenberghs and Verbeke (2005), Rotnitzky and Jewell (1990)]. We can calculate ν_1 based on estimates of $J(\lambda_0)^{(l)}$ and $I(\lambda_0)^{(l)}$ for each locus. Details are given in Appendix C.

3. Simulation study. We have investigated this approach for estimating the relative rate of recombination to mutation using simulation. We used `simMLST` [Didelot, Lawson and Falush (2009)] to simulate MLST data sets, under a standard neutral model for evolution. This involves a model for recombination where the tract length of a recombination event is geometrically distributed. As a default scenario we used parameter values that are appropriate for MLST data for a range of bacteria. This involves data at 7 loci, a population scaled mutation rate, θ, of 100 across the 7 loci, $\lambda = 1$, and mean recombination tract lengths that are 5 times the length of the gene fragments used for each MLST locus. We then considered performance of the method as we varied θ, λ and the number of loci. Note that for `simMLST` recombination rate is defined in terms of the rate at which any loci are affected by recombination, and hence is equal to the product of λ and θ.

3.1. *Estimating* λ. For each scenario we present results from analyzing each locus individually (denoted *Individual*) and results for a combined analysis of data at all loci under the assumption of a common λ value (denoted *Joint*). Results are averaged across 100 simulated data sets for each scenario, and further averaged across loci for the individual analysis. We present results in terms of estimating λ, the relative rate of recombination to mutation. In all cases we look at the bias of the estimates, their root mean square error and the coverage of putative 95% confidence intervals. For each batch of simulations we also present the average number of STs and the number of SLVs (across all loci) per data set. When estimating the distribution of the number of nucleotides introduced at a recombination event (see Appendix A), we assumed that with probability 0.8 a recombination event changed the complete DNA sequence in a region.

Table 2 shows results as we vary the number of isolates in our sample and the number of loci. First consider the individual analysis. Increasing N or L makes only a small impact on the quality of inference. Note that the root mean square error does not decrease much as L increases, because we are independently estimating a value of λ for each locus and the number of SLV pairs per locus is actually reducing. Coverage values are close to their nominal level.

Combining information across loci gives more accurate estimates in all cases as measured by root mean square error; however, coverage proportions drop substantially below the nominal level in many cases. We believe this is because of a slight bias in our estimates, due to the approximate model that we are fitting. The impact of this bias is seen more strongly when the uncertainty of the estimates is lower, such as when combining information across multiple loci. For our simulations we can test whether this is the case, because we are able to "cheat" and remove the bias and then see if the resulting confidence intervals have appropriate coverage

TABLE 2

Results for estimating the relative rate of recombination to mutation for a different number of samples, N, and loci, L. In each case we give the mean number of STs and SLVs per data set, the bias and root mean square error of estimating λ, and the coverage for putative 95% confidence intervals for λ. All simulations had $\lambda = 1$. For the simulations with $L = 7$ we fixed $\theta = 100$ across the 7 loci. For simulations with other numbers of loci we fixed $\theta = 14$ per locus. Individual *gives the results for analyzing a single-locus*, joint *for combining inferences across all loci*

			Individual			Joint		
	STs	**SLVs**	**Bias**	**RMSE**	**Coverage**	**Bias**	**RMSE**	**Coverage**
N					$L = 7$			
5000	629	600	−0.06	0.3	0.92	−0.1	0.15	0.80
10,000	758	797	−0.04	0.26	0.94	−0.06	0.12	0.85
20,000	884	1000	−0.03	0.25	0.92	−0.06	0.11	0.87
40,000	1013	1231	−0.01	0.23	0.93	−0.03	0.09	0.91
L					$N = 10{,}000$			
5	574	730	0	0.29	0.92	−0.03	0.12	0.94
10	973	891	−0.05	0.27	0.94	−0.08	0.11	0.83
20	1591	1196	−0.07	0.31	0.92	−0.11	0.14	0.51

probabilities. We remove the bias by multiplying our estimate of λ by an appropriate constant, chosen so that this new estimator is unbiased. If we do this, coverage values for the joint analysis for all scenarios we considered were in the range 93% to 97%, which is consistent with the nominal significance level, suggesting that we are assessing correctly the uncertainty in our estimators.

We also looked at how the parameters for the mutation and recombination rate affected performance. These are given in Table 3. We see that for a fixed mutation rate, we get similar performance for a range of λ values. Note that while the bias and root mean square error is increasing as λ increases, this is because we are estimating a larger value: the relative size of bias and error remains roughly constant. As we vary θ we notice that performance gets better as θ increases. The large root mean square error values for small θ are caused by estimating a ratio, and the distribution of the estimator in these cases is highly skewed with occasional large estimates for λ. Again, we get more accurate estimates when we combine information across loci, but the coverage values drop noticeably below their nominal level. As mentioned above, this is due to slight bias in the model, which has greater impact for the joint analysis.

3.2. *Testing for variation in* λ. We further looked at both type I error rate and power for testing for a difference in λ across loci. For all the scenarios presented in Section 3.1 we implemented our test for detecting a difference in λ values across loci. We implemented the test at the 95% significance level. These scenarios all correspond to the case where there is a common λ value. The average type I error

TABLE 3

Results for estimating the relative rate of recombination to mutation for different mutation and recombination rates. We use $L = 7$ and $N = 10{,}000$ in all cases. In each case we give the mean number of STs and SLVs per data set, the bias and root mean square error of estimating λ, and the coverage for putative 95% confidence intervals for λ. Individual gives the results for analyzing a single-locus, joint for combining inferences across all loci

			Individual			Joint		
	STs	**SLVs**	**Bias**	**RMSE**	**Coverage**	**Bias**	**RMSE**	**Coverage**
λ				$\theta = 100$				
0.2	524	583	0	0.08	0.94	0	0.04	0.86
0.5	612	665	−0.02	0.15	0.95	−0.03	0.07	0.90
2	1007	1010	−0.12	0.49	0.93	−0.18	0.26	0.80
5	1648	1480	−0.39	1.4	0.9	−0.59	0.73	0.72
θ				$\lambda = 1$				
20	188	236	1.92	9.22	0.93	−0.05	0.29	0.93
50	426	485	−0.03	0.45	0.92	−0.09	0.17	0.84
200	1273	1218	−0.05	0.19	0.93	−0.06	0.1	0.82
500	2374	1908	−0.03	0.14	0.94	−0.04	0.07	0.87

rate across these scenarios was 7%, with a range of 3% to 12%. There was no obvious pattern to which scenarios had higher, or lower, type I error rates, and the range of values observed across the scenarios is consistent with the random fluctuations one would expect if the type I error rate was the same in all cases.

We then investigated the power of the test. We simulated data at 7 loci, with the population-scaled rate at which each locus is affected by recombination being 20, but with different mutation rates per locus. We tried two sets of scenarios: (a) where one locus had $\theta = 20$ and the others had $\theta = 20/C$; and (b) where one locus had $\theta = 20/C$ and all others had $\theta = 20$. The first scenario is where most loci have $\lambda = C$, but one locus has $\lambda = 1$; for the second scenario this is reversed. We repeated this for $C = 2, 3, 4$, reflecting different levels of variation in λ across the loci. Such data sets can be simulated using `simMLST` with $\theta = 20$ per locus and $\lambda = 1$, and then using thinning (removing mutations at a proportion of sites) to reduce the mutation rate at some loci. It is not possible to use `simMLST` to generate data with the same θ but different ρ per locus.

Results are given in Table 4. The results suggest large power for detecting variation in λ of a factor of three or more in scenario (a). There is less power under scenario (b), as there are fewer mutations at the one locus for which λ is different to the others, and this makes it harder to detect the difference in λ.

3.3. *Comparison with a bootstrap approach.* Finally, we give a simple comparison with a parametric bootstrap approach to calculating confidence intervals. The aim here is to give some insight into the challenges and issues relating to

TABLE 4

Power of test for detecting variation in λ across loci. In all cases we simulated data of sample size 10,000 *at* 7 *loci, with $\rho = 20$ per locus. In scenario* (a) *we have one locus with $\theta = 20$ and all others with $\theta = 20/C$; in scenario* (b) *we have one locus with $\theta = 20/C$ and all other scenarios with $\theta = 20$*

C	(a)	(b)
2	0.46	0.33
3	0.93	0.64
4	0.97	0.71

the use of the bootstrap, over and above the extra computational cost it incurs. To mimic issues that occur in real data, we will use a different model to simulate the data than that which we assume when implementing the bootstrap. Our simulated data was generated by `simMLST` under a model where there had been recent population growth. If we define the population size, N, say, to be the size of population prior to the growth, then our model assumes a step change in the population size from N to $10N$ at a time 0.1 in the past. The population-scaled mutation rate is 100 across the 7 loci, and $\lambda = 1$. The effect of this model is to reduce the relative rate of coalescent to mutation and recombination for the larger population size, and hence increase the overall number of mutation and recombination events near the tips of the clonal frame. We simulated 100 data sets, each with sample size of 1000 isolates. A summary of the results from analyzing these data sets using the composite likelihood method is shown in Table 5. The observed type I error rate for the test of a common λ across loci at 95% significance level was 0.03.

To implement a parametric bootstrap, we then simulated data from a constant population size model. We fixed the population-scaled mutation rate to be equal

TABLE 5

Results for estimating relative rate of recombination to mutation for a population growth scenario. We give coverage for three types of putative 95% *confidence intervals: those from asymptotic theory for composite likelihood, and two parametric bootstrap approaches. The latter differ in whether they simulate data sets that match the true data in terms of the number of STs or the number of SLVs. Results are shown for both analyzing loci individually and a joint analysis*

	Population growth (Average 810 STs 173 SLVs)				
	Composite likelihood			Bootstrap	
	Bias	RMSE	Coverage	Coverage (ST)	Coverage (SLV)
Individual	0.03	0.56	0.95	0.69	0.98
Joint	−0.08	0.17	0.93	0.65	0.99

to the average number of the difference between a pair of isolates, and set $\lambda = 1$. Data sets simulated from a constant population size model have different patterns in terms of the ratio of SLVs to STs that are observed for the population growth model. Thus, we considered two approaches to simulating data for the parametric bootstrap. For each "real" data set we analyzed we simulated 100 data sets under the constant population size model. Our first approach simulated each of these 100 data sets to have the same number of STs as the real data set, while the second approach matched in terms of the number of SLVs. The use of the parametric bootstrap added considerably to the cost of analyzing the data. The composite likelihood approach takes a matter of seconds to run, as compared to of the order of 10 hours to simulate 100 SLV data sets.

To construct our parametric bootstrap confidence interval, we used the approach of Yu et al. (2012). This is based on noting that the sampling distribution of an estimate of $\lambda/(1+\lambda)$ is approximately normal. We used the 100 data sets simulated under the parametric bootstrap to estimate the variance of this normal distribution. This enables us to produce putative 95% symmetric confidence intervals for $\lambda/(1+\lambda)$, which can then be transformed to confidence intervals for λ.

The observed coverage for each of these two methods for constructing bootstrap confidence intervals is shown in Table 5. We observe that choosing the size of the data sets simulated by the parametric bootstrap through matching the number of STs leads to much smaller confidence intervals than when we match on the number of SLVs. The coverage for the intervals when we match on STs is substantially lower than the putative 95% confidence level. In this example matching on SLVs gives much better results, though the coverage results suggest that the confidence intervals produced are slightly too conservative.

The main message from these simulations is that when performing a parametric bootstrap there can be issues if features of the simulated data sets do not match the real data set when these features affect the amount of information the data has about the parameter of interest. These issues disappear if we are able to simulate from a model which is a very good approximation to real life, but for bacterial data this is rarely the case.

4. Application to bacterial MLST data. We applied our composite likelihood method to detect variation in λ across loci and estimate λ for a range of bacteria. We used MLST data downloaded from http://pubmlst.org/. In each case we had data at 7 loci. The bacteria we considered, together with the number of SLV pairs we obtained, were as follows: *Bacillus cereus* (281 SLV pairs); *Enterococcus faecium* (481); *Haemophilus influenzae* (977); *Klebsiella pneumoniae* (404); *Staphylococcus aureus* (7892); *Streptococcus uberis* (356); *Campylobacter jejuni* (7417); and *Campylobacter coli* (1842).

We found evidence for variation of λ across loci (p-values less than 0.01) in 3 bacteria: *B. cereus* (p-value 2.2×10^{-8}); *E. faecium* (9.4×10^{-6}); and *K. pneumoniae* (4.9×10^{-12}). Estimates of λ for each locus for these bacteria are given in Figure 3, and estimates of a common λ for the other bacteria are given in Table 6.

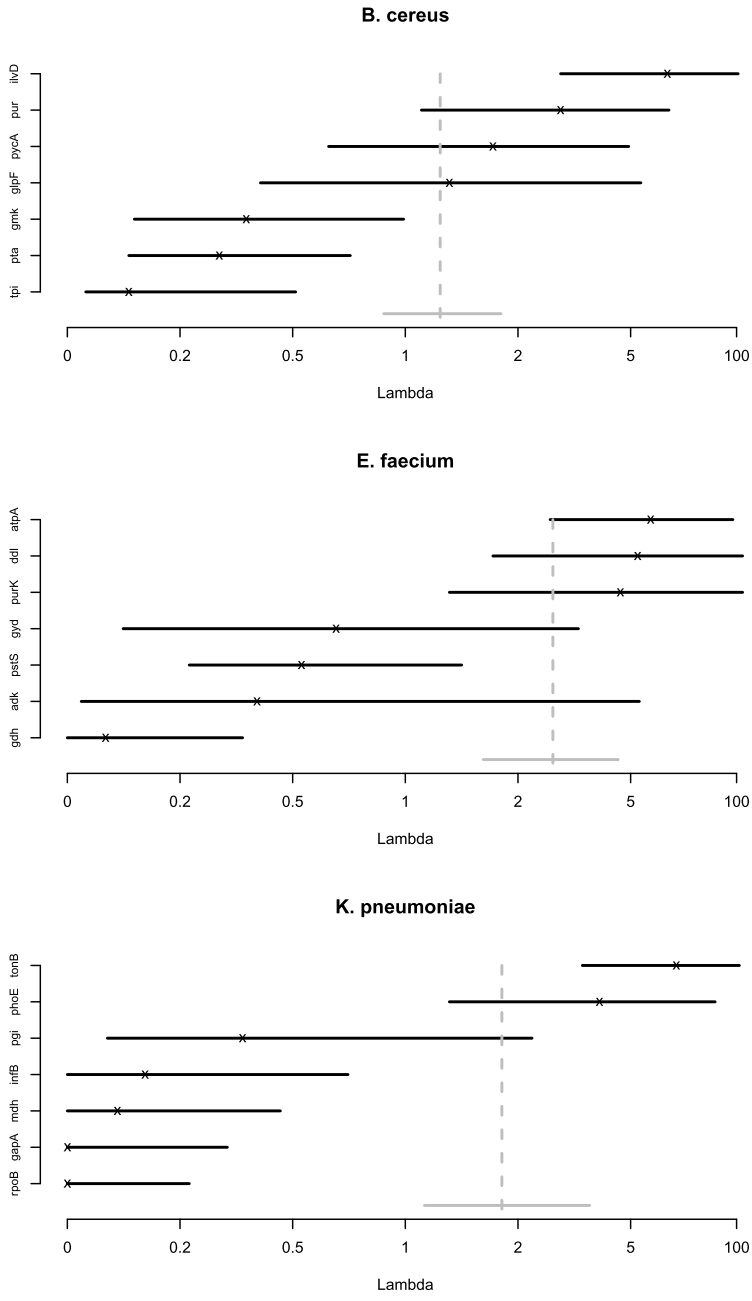

FIG. 3. *Estimates and confidence intervals for* λ *for the three bacteria that showed evidence of variation across loci. For each bacteria we plot the estimate* (*cross*) *and putatitve* 95% *confidence intervals* (*lines*) *for each locus. In gray is the estimate* (*vertical dashed line*) *and* 95% *confidence interval* (*horizontal line*) *under an assumption of a common value of* λ *across loci. We have ordered the loci in terms of the value of the estimate of* λ*, with decreasing estimates as we move down each plot. For clarity we have chosen an* $x/(1+x)$ *scale for the* x*-axis for each plot.*

TABLE 6
Estimate of common λ across MLST loci, together with putative 95% confidence intervals

Bacteria	Estimate of λ	95% CI
H. influenzae	4.9	(3.3, 7.4)
S. aureus	1.4	(0.92, 2.1)
S. uberis	11	(4.8, 180)
C. jejuni	3.4	(2.9, 4.1)
C. coli	0.43	(0.21, 0.88)

The most striking results are the degree of variation we see in estimates of λ for *B. cereus*, *E. Faecium* and *K. pneumoniae*. For each bacteria we have evidence for loci with $\lambda < 1$, and often $\lambda \approx 0$, as well as for loci with λ substantially greater than 1. The estimates we get are consistent with the relative rate of recombination to mutation varying by between one and two orders of magnitude across the loci for each bacteria.

For the five bacteria for which we see no evidence for variation in λ, we see that the relative rate of recombination to mutation varies noticeably across the bacteria. The order of bacteria from the one with highest to lowest λ is *S. uberis*, *H. influenzae*, *C. jejuni*, *S. aureus* and *C. coli*. The estimates of λ for *S. uberis* and *C. coli* differ by a factor of 25.

5. Conclusion. We have presented a way of both estimating the relative rate of recombination to mutation and detecting recombination rate variation from MLST data. The key novelty within the statistical approach we take is to directly model the form of dependence of information from different SLVs. This enables us to correct for the dependence between the contribution each SLV makes to the composite likelihood, and thus get appropriate measures of uncertainty in the estimates of parameters we get from maximizing the composite likelihood, and also to test for variation in recombination rate across loci. While composite likelihood methods are extensively used within genetics [see, e.g., Larribe and Fearnhead (2011) and references therein], to date they have been primarily used to get point estimates of parameters. Our results suggest that, with appropriate modeling of the dependence structure, it should be possible to extend these earlier methods to obtain both point estimates and confidence intervals for parameters of interest.

An example of why this is important can be seen by a previous analysis of MLST data in bacteria. Recombination rates in *B. cereus* at MLST loci were estimated in Pérez-Losada et al. (2006), using an alternative composite likelihood approach [Hudson (2001), McVean, Awadalla and Fearnhead (2002)]. They observed estimates of the recombination rate varying by a factor of nearly 50 across the loci; however, they were unable to calculate confidence intervals for their recombination rate estimates. As a result, it was impossible to conclude whether this

variation is due to variability in the estimator or whether it resulted from true differences in recombination rates across the loci. By comparison, our analysis gives strong evidence for variation in λ across these loci.

Recently MLST data are increasingly being replaced by full-genome data. The methods developed here can still be applied to such data, by choosing a set of L loci and summarizing the full-genome data in terms of SLV pairs and the nucleotide differences of each pair. To be viable, such an approach would need full-genome data from a substantial number of isolates in order to produce sufficient SLV pairs. Summarizing the data in such a way would clearly lead to a substantial loss of information, but would be a simple and quick approach to performing an initial analysis of data as compared to methods that try and analyze the full sequence data [e.g., Didelot et al. (2010)]. As pointed out by a reviewer, the flexibility over the choice of loci would give the possibility of using a nonparametric bootstrap to assess uncertainty in estimates. If interest is in the ratio of recombination to mutation at a given loci, we can make different choices for the other $L - 1$ loci. Each choice would give a different set of SLVs, and hence a different estimate. The variability of these estimates could be used to measure the degree of uncertainty in the final estimate we make.

The results from our application to data from eight bacteria species are in line with results in Vos and Didelot (2009). While that paper estimated a different measure of recombination to mutation, looking at the probability of a nucleotide change as due to recombination rather than mutation, the ordering of bacteria species from less to more recombinant is broadly similar. The more striking results, though, relate to strong evidence of rate variation in the rate of recombination to mutation in three of the bacteria. This is part of the growing evidence for substantial recombination rate variation, for example, in Didelot et al. (2010), who also found evidence of rate variation in *B. cereus*, and Guy et al. (2012), who observed 3 orders of magnitude of recombination rate variation in *Bartonella henselae*. More indirect evidence for rate variation also comes from the variation in recombination rates for closely related bacterial species and substantial differences in estimates of recombination rates from different studies of a given bacterial species [see Didelot and Maiden (2010) for more details].

The reasons behind substantial variation in the relative rate of recombination to mutation are currently unclear. One explanation is that estimated recombination rates are higher within regions under positive selection [Vos (2009)]. The argument is that we are only likely to see recombination events that add beneficial or remove deleterious mutations. The selective advantage of such recombination events over mutation will be largest within genes for which selection is strongest. In our study we see large variation in recombination rates among housekeeping genes, genes we would expect to all be under strong selective pressure. This includes variation in recombination rates between genes with similar function: for example, both *pycA* and *tpi* in *B. cereus* are genes involved in gluconeogenesis, yet their estimates

of λ differ by an order of magnitude. This suggests that there are other factors responsible for the variation that we observe.

The MLST data analyzed in Section 4, together with R code implementing the composite likelihood method presented in this paper, are available from http://www.maths.lancs.ac.uk/~fearnhea/SLV.zip.

APPENDIX A: ESTIMATING $\Pr_\lambda(X = x|\text{SLV}, A^c)$

Our approach to approximating $\Pr_\lambda(X = x|\text{SLV}, A^c)$ for a given locus is to use a Monte Carlo estimate of the probability of x nucleotide differences being imported at a single recombination event. This simple idea is based upon the fact that for an SLV pair we expect the isolates to have a recent common ancestor, and hence the number of mutation/recombination events to be one with a high probability. It also simplifies computations in that this approximation is independent of λ, and hence can be calculated and stored once. It is possible to extend the following Monte Carlo setup to allow for possible multiple events at the locus, but to do this correctly would involve making the approximation of $\Pr_\lambda(X = x|\text{SLV}, A^c)$ depend on λ.

We assume we have a sample of K isolates, and for each pair (i, j) know the number of nucleotide differences at locus l between isolates i and j, denoted x_{ij}. Let m be the number of bases at the SLV locus. Assume with probability p_a the region of the recombination event will include all m bases of the locus. Fix the Monte Carlo sample size, M:

(1) Set $n_i = 0$ for $i = 0, \ldots, m$.
(2) Repeat M times:
   (a) Sample i and j independently from $\{1, 2, \ldots, K\}$.
   (b) With probability p_a set $x = x_{ij}$; otherwise sample u, a realization of a standard uniform random variable, and x the realization of a Binomial random variable with parameters x_{ij} and u.
   (c) Set $n_x = n_x + 1$.
(3) Calculate the approximation

$$\Pr_\lambda(X = x|\text{SLV}, A^c) \approx \frac{n_x + 1}{M + m - n_0}.$$

In step (2b) we have used a simple mechanism for simulating the number of changes due to a recombination event for which a breakpoint lies within the locus. This involves simulating u, the proportion of the locus that is affected by the recombination event, and then, conditional on this, how many nucleotide differences the recombination event introduces.

The final approximation used in part (3) is chosen so that all probabilities are non 0, to allow for the possibility of a value x for the number of nucleotide differences observed for an SLV pair that was not simulated. We subtract n_0 from the denominator, as we are conditioning on there being an SLV pair at locus l, in

which case $x \neq 0$. We repeat this procedure to get a different distribution for the number of nucleotide differences for each locus.

APPENDIX B: ESTIMATING α

Assume we have partitioned the STs in G groups of size $n_1, \ldots, n_G$, and for each SLV pair we have the value of the score function at $\hat{\lambda}$. To estimate the within-group correlation of the score statistics, we will model the scores as being Gaussian, with independence across groups. Within group g we will view the scores as realization of a vector random-variable $V = (u(\lambda_0; X_1), \ldots, u(\lambda_0; X_k))$, where $k = n_g(n_g - 1)/2$. We model $V \sim \mathrm{MVN}(\mathbf{0}, \Sigma)$, and Σ is a $k \times k$ variance–covariance matrix for $i = 1, \ldots, k$, with $\Sigma_{ii} = \sigma^2$; and for $i, j = 1, \ldots, k$ with $i \neq j$, $\Sigma_{ij} = \alpha\sigma^2$. If we denote the data for this group as $v = (u(\hat{\lambda}; x_1), \ldots, u(\hat{\lambda}; x_k))$, then the likelihood for the group is $l(\alpha, \sigma; v) = -0.5 \log\det(\Sigma) - \frac{1}{2} v \Sigma^{-1} v^T$. Using Sylvester's theorem,

$$\det(\Sigma) = \sigma^{2k}(1-\alpha)^k\left(\frac{1+(k-1)\alpha}{1-\alpha}\right) \quad \text{and} \quad \Sigma^{-1} = a_k I_k + b_k 1_k,$$

where I_k is the identity matrix, 1_k is a $k \times k$ matrix of 1's, $a_k = 1/(1-\alpha)$ and $b_k = -\alpha/(1-\alpha)[1+(n-1)\alpha]$. Thus,

$$l(\alpha, \sigma; v) = -\frac{k}{2}\log(\sigma^2[1-\alpha]) - \frac{1}{n}\log\left(\frac{1+(k-1)\alpha}{1-\alpha}\right) - \frac{1}{2}\left(a_k \sum_{i=1}^{k} u(\hat{\lambda}; x_i)^2 + b_k\left[\sum_{i=1}^{k} u(\hat{\lambda}; x_i)\right]^2\right).$$

If we denote the data in group g by $x_1^{(g)}, \ldots, x_{k_g}^{(g)}$ where $k_g = n_g(n_g-1)/2$, then we get a likelihood

$$\sum_{g=1}^{G}\left\{-\frac{k_g}{2}\log(\sigma^2[1-\alpha]) - \frac{1}{n}\log\left(\frac{1+(k_g-1)\alpha}{1-\alpha}\right) - \frac{1}{2}\left(a_{k_g}\sum_{i=1}^{k_g} u(\hat{\lambda}; x_i^{(g)})^2 + b_{k_g}\left[\sum_{i=1}^{k} u(\hat{\lambda}; x_i^{(g)})\right]^2\right)\right\},$$

which we maximize numerically over $\sigma > 0$ and $\alpha \in [0, 1)$ to get an estimate of α. In practice, we use a common value of α for all loci, obtained by averaging the locus-specific estimates.

APPENDIX C: ESTIMATING ν_1

Consider a model for data at L loci. The general model will have parameter vector $(\lambda_1, \ldots, \lambda_L)$ for the value of the rate of recombination to mutation for each of

the L loci. Under our assumption of independence across loci, our joint composite log-likelihood is $\mathrm{Cl}^*(\lambda_1, \ldots, \lambda_L) = \sum_{l=1}^{L} \mathrm{Cl}^{(l)}(\lambda_l)$, the sum of the composite log-likelihoods for each locus.

Assume that there is a common λ value for all loci. Let λ_0 denote the true common value. Further, to simplify notation, let $J^{(l)} = J^{(l)}(\lambda_0)$ and $I^{(l)} = I^{(l)}(\lambda_0)$ be the value of J and I at locus l evaluated at this true common value. Then the J and I matrices associated with our joint composite log-likelihood will be diagonal with entries $(J^{(1)}, \ldots, J^{(L)})$ and $(I^{(1)}, \ldots, I^{(L)})$, respectively.

We are interested in a test for whether there is a common λ value for all loci. To do this, we can introduce a reparameterization to $(\phi_1, \ldots, \phi_L)$, where $\phi_1 = \lambda_1$, and for $l = 2, \ldots, L$, $\phi_l = \lambda_l - \lambda_1$. So a common λ value is equivalent to $\phi_l = 0$ for $l = 2, \ldots, L$. Let

$$\mathrm{Cl}_\phi(\phi_1, \ldots, \phi_L) = \mathrm{Cl}^*(\phi_1, \phi_2 + \phi_1, \ldots, \phi_L + \phi_1)$$

be the composite log-likelihood under this parameterization. The corresponding J and I matrices will be denoted by J_ϕ and I_ϕ. The diagonal, first row and column of J_ϕ are $(\sum_{l=1}^{L} J^{(l)}, J^{(2)}, J^{(3)}, \ldots, J^{(L)})$, with all other entries being 0; and I_ϕ depends on $I^{(1)}, \ldots, I^{(L)}$ in a similar way.

The likelihood ratio statistic for testing $\phi_l = 0$ for $l = 2, \ldots, L$ is given by

$$\begin{aligned} LR^* &= 2\big[\max \mathrm{Cl}_\phi(\phi_1, \ldots, \phi_L) - \max \mathrm{Cl}_\phi(\phi_1, 0, \ldots, 0)\big] \\ &= 2\left[\sum_{l=1}^{L} \max \mathrm{Cl}^{(l)}(\lambda) - \max \sum_{l=1}^{L} \mathrm{Cl}^{(l)}(\lambda)\right]. \end{aligned}$$

Define H and G to be $(L-1) \times (L-1)$ matrices obtained from removing the first row and column from I_ϕ^{-1} and $(I_\phi J_\phi^{-1} I_\phi)^{-1}$, respectively. Let η_i be the eigenvalues of the matrix $H^{-1}G$. Then if $\phi_l = 0$ for $l = 2, \ldots, L$, the asymptotic distribution of LR^* is $\sum_{i=1}^{L-1} \eta_i Z_i^2$, where $Z_1, \ldots, Z_{L-1}$ are independent standard normal random variables [see Kent (1982), Varin, Reid and Firth (2011)].

We approximate this distribution by scaling LR^* to have the same mean as a chi-squared distribution with $L-1$ degree of freedom, χ^2_{L-1} [Rotnitzky and Jewell (1990)]. Thus, we define $\nu_1 = \sum_{i=1}^{L-1} \eta_i/(L-1)$, set $LR = (1/\nu_1)LR^*$, and approximate the distribution of LR by a χ^2_{L-1} distribution. Higher order approximations are possible [Varin (2008)], but we did not find them to be more accurate in practice.

Acknowledgments. This publication made use of the PubMLST website (http://pubmlst.org/) developed by Keith Jolley [Jolley and Maiden (2010)] and sited at the University of Oxford. The development of this site has been funded by the Wellcome Trust.

REFERENCES

DIDELOT, X. and FALUSH, D. (2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175** 1251–1266.

DIDELOT, X., LAWSON, D. and FALUSH, D. (2009). SimMLST: Simulation of multi-locus sequence typing data under a neutral model. *Bioinformatics* **25** 1442–1444.

DIDELOT, X. and MAIDEN, M. C. J. (2010). Impact of recombination on bacterial evolution. *Trends in Microbiology* **18** 315–322.

DIDELOT, X., LAWSON, D., DARLING, A. and FALUSH, D. (2010). Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* **186** 1435–1449.

DONNELLY, P. and TAVARÉ, S. (1995). Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29** 401–421.

FEIL, E. J., MAIDEN, M. C. J., ACHTMAN, M. and SPRATT, B. G. (1999). The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Molecular Biology and Evolution* **16** 1496–1502.

FEIL, E. J., SMITH, J. M., ENRIGHT, M. C. and SPRATT, B. G. (2000). Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. *Genetics* **154** 1439–1450.

FRASER, C., HANAGE, W. P. and SPRATT, B. G. (2007). Recombination and the nature of bacterial speciation. *Science* **315** 476–480.

GRIFFITHS, R. C. and MARJORAM, P. (1997). An ancestral recombination graph. In *Progress in Population Genetics and Human Evolution* (*Minneapolis*, *MN*, 1994). *IMA Vol. Math. Appl.* **87** 257–270. Springer, New York. MR1493031

GUY, L., NYSTEDT, B., SUN, Y., NASLUND, K., BERGLUND, E. and ANDERSSON, S. G. (2012). A genome-wide study of recombination rate variation in *Bartonella henselae*. *BMC Evolutionary Biology* **12** 65.

HUDSON, R. R. (2001). Two-locus sampling distributions and their application. *Genetics* **159** 1805–1817.

JOLLEY, K. A. and MAIDEN, M. C. J. (2010). Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11 595**.

KENT, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika* **69** 19–27. MR0655667

LARRIBE, F. and FEARNHEAD, P. (2011). On composite likelihoods in statistical genetics. *Statist. Sinica* **21** 43–69. MR2796853

LOW, K. B. and PORTER, D. D. (1978). Modes of gene transfer and recombination in bacteria. *Annu. Rev. Genet.* **12** 249–287.

MAIDEN, M. C. J., BYGRAVES, J. A., FEIL, E., MORELLI, G., RUSSELL, J. E., URWIN, R., ZHANG, Q., ZHOU, J., ZURTH, K., CAUGANT, D. A., FEAVERS, I. M., ACHTMAN, M. and SPRATT, B. G. (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. USA* **95** 3140–3145.

MCVEAN, G. A. T., AWADALLA, P. and FEARNHEAD, P. (2002). A coalescent method for detecting recombination from gene sequences. *Genetics* **160** 1231–1241.

MILKMAN, R. and BRIDGES, M. M. (1990). Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. *Genetics* **126** 505–517.

MOLENBERGHS, G. and VERBEKE, G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York. MR2171048

PÉREZ-LOSADA, M., BROWNE, E. B., MADSEN, A., WIRTH, T., VISCIDI, R. P. and CRANDALL, K. A. (2006). Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infection, Genetics and Evolution* **6** 97–112.

ROTNITZKY, A. and JEWELL, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77** 485–497. MR1087838

SHEPPARD, S. K., MCCARTHY, N. D., FALUSH, D. and MAIDEN, M. C. J. (2008). Convergence of Campylobacter species: Implications for bacterial evolution. *Science* **320** 237–239.

SPRATT, B. G., HANAGE, W. P. and FEIL, E. J. (2001). The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Current Opinion in Microbiology* **4** 602–606.

VARIN, C. (2008). On composite marginal likelihoods. *AStA Adv. Stat. Anal.* **92** 1–28. MR2414624

VARIN, C., REID, N. and FIRTH, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21** 5–42. MR2796852

VOS, M. (2009). Why do bacteria engage in homologous recombination? *Trends Microbiol.* **17** 226–232.

VOS, M. and DIDELOT, X. (2009). A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* **3** 199–208.

WAKELEY, J. (2007). *Coalescent Theory: An Introduction*. Roberts and Company, Denver, CO.

YU, S., FEARNHEAD, P., HOLLAND, B. R., BIGGS, P., MAIDEN, M. and FRENCH, N. (2012). Estimating the relative roles of recombination and point mutation in the generation of single locus variants in *Campylobacter jejuni* and *Campylobacter coli*. *J. Mol. Evol.* **74** 273–280.

P. FEARNHEAD
DEPARTMENT OF MATHEMATICS AND STATISTICS
FYLDE COLLEGE
LANCASTER UNIVERSITY
LANCASTER LA1 4YF
UNITED KINGDOM
E-MAIL: p.fearnhead@lancaster.ac.uk

S. YU
P. BIGGS
N. FRENCH
INFECTIOUS DISEASE RESEARCH CENTRE
INSTITUTE OF VETERINARY, ANIMAL AND BIOMEDICAL SCIENCES
PRIVATE BAG 11 222
MASSEY UNIVERSITY
PALMERSTON NORTH 4442
NEW ZEALAND
E-MAIL: S.Yu1@massey.ac.nz
P.Biggs@massey.ac.nz
N.P.French@massey.ac.nz

B. HOLLAND
SCHOOL OF MATHEMATICS AND PHYSICS
UNIVERSITY OF TASMANIA
HOBART
TASMANIA 7001
AUSTRALIA
E-MAIL: Barbara.Holland@utas.edu.au