Web-based Supplementary Materials for "Estimating the transmission dynamics of *Streptococcus Pneumoniae* from strain prevalence data" by E. Numminen, L. Cheng, M. Gyllenberg and J. Corander.

0.1 The Adaptive Enhanced ABC

In sequential ABC methods, a sequence of target distributions is defined along with a sequence of decreasing tolerance values, as $f_t = f(\boldsymbol{\varphi}|\rho(\boldsymbol{S}(\boldsymbol{D}), \boldsymbol{S}(\boldsymbol{D'})) < \boldsymbol{\varepsilon^t})$. Starting from the largest tolerance and proceeding in order, ABC-sampling is performed for each ε^t until an approximation of f_t is obtained. When approximating f_1, parameters are proposed from the prior distribution, otherwise f_t is used to construct a proposal distribution for $t+1$. Then, to account for the effect of this sampling distribution, an importance weight is given to each accepted parameter. Finally, when the parameters are sampled according to these weights, they are distributed as in f_{t+1}.

In our approach $\boldsymbol{\varepsilon^t} = (\varepsilon_1^t, \varepsilon_2^t, \varepsilon_3^t, \varepsilon_4^t)$, and for each $k = 1, ..., 4$, we require $\varepsilon_k^t \geq \varepsilon_k^{t+1}$. Rather than defining the sequence of tolerances in advance, we set $\boldsymbol{\varepsilon^{t+1}}$ dynamically, based on the approximation of the distribution of the discrepancies $\{d_k\}$ obtained with $\boldsymbol{\varepsilon^t}$. We construct empirical distribution functions for the discrepancies of each simulated summary, denoted with G_k^t, and the joint empirical distribution function for all the four discrepancies, denoted with G^t. We adjust the tolerance for generation $t+1$ by numerically searching for such values of ε_j and q, so that the following holds:

$$G^t(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4) = a \quad \text{and } G_j^t(\varepsilon_j) = q, \quad \text{for all } j = 1, ..., 4 \tag{1}$$

The resulting vector of tolerances $\boldsymbol{\varepsilon}$ is taken as tolerance for generation $t+1$. Now $\boldsymbol{\varepsilon}$ is such that a proportion a of the accepted parameters in generation t was within $\boldsymbol{\varepsilon}$, and ε_j is the same quantile of each the sets of discrepancies $\{d_j\}$, corresponding to the accepted parameters of the previous generation. This approach is similar to that proposed by Del Moral, Doucet and Jasra (2011), except that they consider the tolerance as one-dimensional, whereas having a vector of tolerances complicates the problem because $G^t(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4) = a$ does not have a unique solution.

Prior distributions were used as proposal distributions in the first ABC-generation. In the following generations, the accepted parameters of the previous generation are sampled according to their importance weights, and to improve the search on the parameter space, each parameter $\boldsymbol{\varphi}$ is then perturbed randomly with a perturbation kernel $K(\boldsymbol{\varphi} \mid \bullet)$ to $\boldsymbol{\varphi'}$, and a dataset is generated from $\boldsymbol{D'} \sim \Phi(\bullet, \boldsymbol{\varphi'})$. For the perturbation kernel $K(\boldsymbol{\varphi}, \bullet)$, we use the multivariate normal distribution with mean $\boldsymbol{\varphi}$ and the covariance matrix estimated from the accepted parameters in the previous generation. See Filippi et al. (2011) for a discussion of such adaptive tuning of the perturbation kernels. Algorithm 1 below describes the computation in more detail, where $\mu(\boldsymbol{\varphi})$ is used

to denote the prior distribution of the parameters. The algorithm is similar to

Algorithm 1 The adaptive enhanced ABC-PRC algorithm

1: **Initialization**: Set ABC-generation indicator to $t = 1$.
2: Set initial tolerance $\boldsymbol{\varepsilon}^1$ and sample size n.
3: **for** $j = 1$ **to** n **do**
4: **repeat**
5: Sample $\boldsymbol{\varphi}^t(j) \sim \mu(\boldsymbol{\varphi})$ and $\boldsymbol{D}^t \sim \Phi(\bullet|\boldsymbol{\varphi}_t(j))$
6: Determine $\boldsymbol{d}^t(j)$,
7: **until** $d^t(j)$ is within $\boldsymbol{\varepsilon^t}$.
8: **for** $j = 1$ **to** n **do**
9: Set $w^t(j) = \frac{1}{n}$.
10: **Set ABC-generation indicator to** $t = t + 1$**.**
11: Initialize the tolerance $\boldsymbol{\varepsilon}^t$ and $K(\boldsymbol{\varphi}, \bullet)$, based on $\{\boldsymbol{\varphi}^{t-1}(k),\ \boldsymbol{d}^{t-1}(k),\ w^{t-1}(k) \mid k = 1, ..., n\}$.
12: **for** $j = 1$ **to** n **do**
13: **repeat**
14: Sample $\varphi*$ from $\{\boldsymbol{\varphi}^{t-1}(k) \mid k = 1, ..., n\}$ according to $\{w^{t-1}(k) \mid k = 1, ..., n\}$,
15: Sample $\boldsymbol{\varphi}^t(j) \sim K(\boldsymbol{\varphi}*, \bullet)$ and $D^t \sim \Phi(\bullet|\boldsymbol{\varphi}^t(j))$
16: Determine $\boldsymbol{d}^t(j)$,
17: **until** $\boldsymbol{d}^t(j)$ is within $\boldsymbol{\varepsilon}^t$.
18: **for** $j = 1$ **to** n **do**
19: Set $w^t(j) \propto \frac{\mu(\boldsymbol{\varphi}^t(j))}{\sum_{i=1}^n w^{t-1}(i) K(\boldsymbol{\varphi}^{t-1}(i), \boldsymbol{\varphi}^t(j))}$, so that $\sum_{j=1}^n w^t(j) = 1$.
20: **Determine whether to continue for a further ABC generation. If the decision is to continue go to step 10.**

that presented by Sisson, Fan and Tanaka (2007), except that the tolerances are adjusted as we have described. We perform sampling for each ABC-generation t until a sample size of $n = 10,000$ accepted parameter values is reached. Tolerance for each generation is adjusted using $a = 0.1$, and the condition 1. For the first generation, training simulations were utilized to set the tolerance according to this criteria. Whether further ABC-generations are needed is determined by comparing the two most recently sampled approximate posterior distributions by manual inspection.

References

Del Moral, P., Doucet, A. and Jasra, A. (2011) An Adaptive Sequential Monte Carlo Method for Approximate Bayesian Computation, *Statistics and Computing* **DOI** 10.1007/s11222-011-9271-y .

Filippi, S., Barnes, C., Cornebise, J. and Stumpf, M.P.H. (2011) On Optimality of Kernels for approximate Bayesian Computation using sequential Monte Carlo *arXiv*:1106.6280v3.

Sisson, SS., Fan, Y. and Tanaka MM (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* **104,** 1760-1765.

Web Figure 1

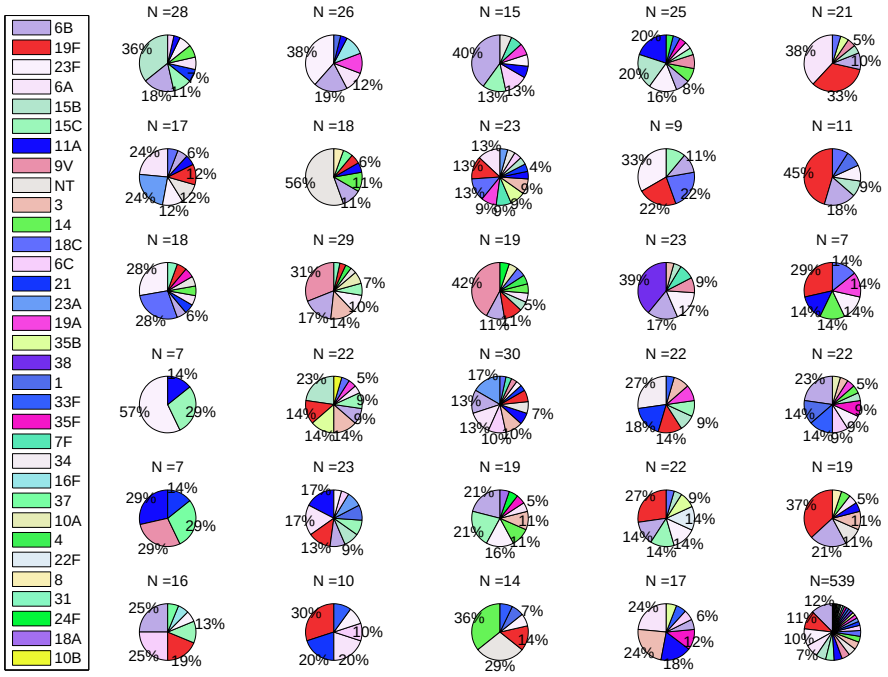

Figure 1: The figure shows the proportions of different strains among all the pneumococcal isolates obtained from each DCC. The total numbers of isolates obtained from the particular DCC are shown above the charts. The last pie chart illustrates the strain distribution among all positive swabs.

Web Figure 2

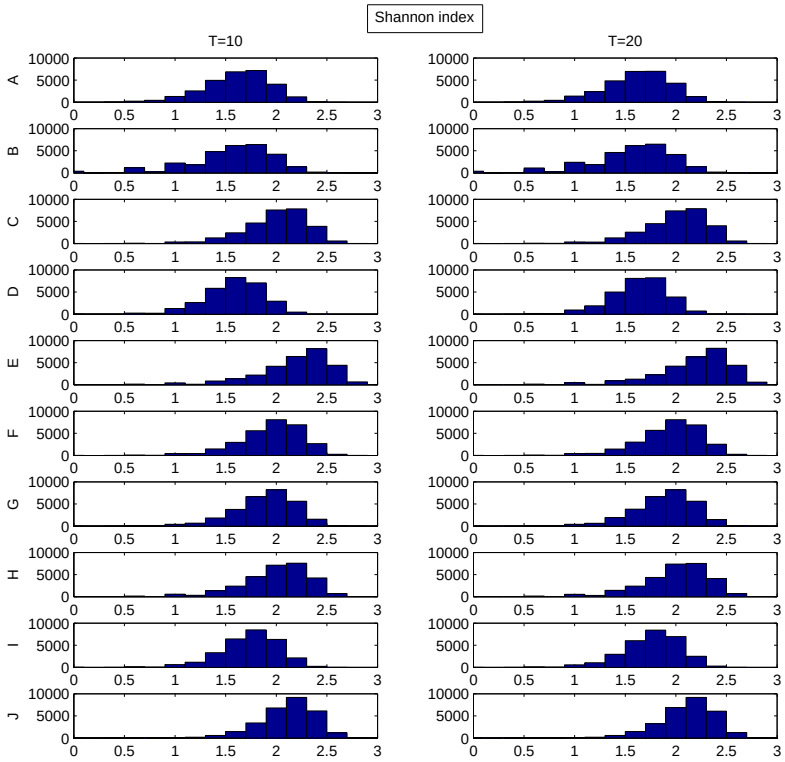

Figure 2: Figure shows the distribution for observations on Shannon index, first column being the simulation results with $T = 10$ and the second with simulation time $T = 20$. Each row corresponds to a different model parameter combination. Parameter combinations corresponding to each row are given in Web Table 1. The similarity of distributions with both simulation times suggest, that $T = 10$ was enough long time for sampling from the stationary distribution of Shannon index. By repeating the comparison for other summaries of the data, we conclude that for all the tested parameter combinations, $T = 10$ is a simulation time yields observations distributed that are likely distributed similar to stationary distribution.

Web Figure 3

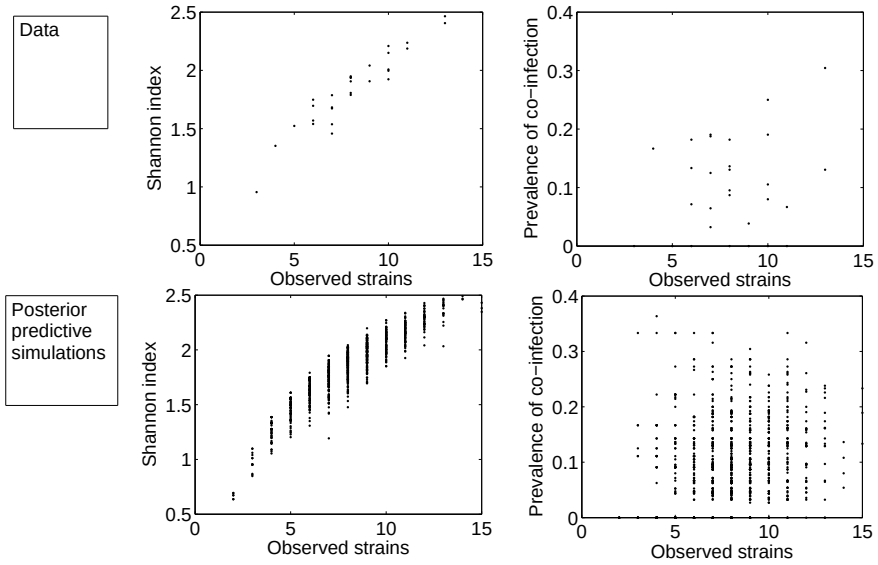

Figure 3: In the first row we show the correlation between two summary pairs as observed in the data, while in the second row we show the correlations predicted by the posterior predictive simulations.

Web Figure 4

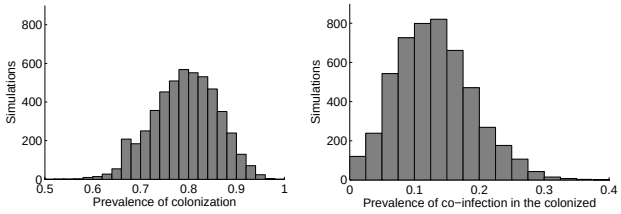

Figure 4: The figure shows the predicted prevalences of colonization and co-infection, in a single DCC consisting of 53 individuals.

Web Table 1

Table 1: The parameters used in simulation of distributions shown in Web Figure 2

	β	Λ	θ
A	10.6000	1.0000	0.0100
B	1.4487	0.2472	0.1000
C	4.7700	1.8866	0.0040
D	2.4733	0.0592	0.7300
E	0.0018	1.7472	0.1800
F	3.0984	0.6238	0.1408
G	5.3710	0.8471	0.0649
H	2.3323	1.2100	0.0169
I	2.0386	0.0846	0.8683
J	7.1338	1.4485	0.0939