

1 **Sequence element enrichment analysis to determine the** 2 **genetic basis of bacterial phenotypes**

3

4 John A. Lees^{1†}, Minna Vehkala^{2†}, Niko Välimäki³, Simon R. Harris¹, Claire
5 Chewapreecha⁴, Nicholas J. Croucher⁵, Pekka Marttinen^{6,7}, Antti Honkela⁸, Julian
6 Parkhill¹, Stephen D. Bentley¹, Jukka Corander^{2*}

7 ¹Pathogen Genomics, Wellcome Trust Sanger Institute, Cambridge, UK

8 ²Department of Mathematics and Statistics, University of Helsinki, Helsinki,
9 Finland

10 ³Department of Medical and Clinical Genetics, Genome-Scale Biology Research
11 Program, University of Helsinki

12 ⁴Department of Medicine, University of Cambridge, Cambridge, UK

13 ⁵Department of Infectious Disease Epidemiology, Imperial College, London, UK

14 ⁶Department of Computer Science, Aalto University, Espoo, Finland

15 ⁷Helsinki Institute of Information Technology HIIT, Department of Computer
16 Science, Aalto University, Espoo, Finland

17 ⁸Helsinki Institute for Information Technology HIIT, Department of Computer
18 Science, University of Helsinki, Helsinki, Finland

19

20 * Corresponding author: jukka.corander@helsinki.fi

21 † These authors contributed equally.

22 **Abstract**

23 Bacterial genomes vary extensively in terms of both gene content and gene
24 sequence – this plasticity hampers the use of traditional SNP-based methods for
25 identifying all genetic associations with phenotypic variation. Here we introduce
26 a computationally scalable and widely applicable statistical method (SEER) for
27 the identification of sequence elements that are significantly enriched in a
28 phenotype of interest. SEER is applicable to even tens of thousands of genomes
29 by counting variable-length k-mers using a distributed string-mining algorithm.
30 Robust options are provided for association analysis that also correct for the
31 clonal population structure of bacteria. Using large collections of genomes of the
32 major human pathogen *Streptococcus pneumoniae*, SEER identifies relevant
33 previously characterised resistance determinants for several antibiotics. We thus
34 demonstrate that our method can answer important biologically and medically
35 relevant questions.
36

37 **Introduction**

38 The rapidly expanding repositories of genomic data for bacteria hold an
39 enormous and yet largely untapped potential for building a more detailed
40 understanding of the evolutionary responses to changing environmental
41 conditions, such as the widespread use of antibiotics and switches between host-
42 niche as farming practices change.

43

44 Genome-wide association studies (GWAS) for bacterial phenotypes have only
45 recently started to appear¹⁻⁵. Use of standard GWAS methods developed
46 originally for human SNP data have been shown to be successfully applicable to
47 core genome mutations in bacteria^{2,3}. However, given the high level of genome
48 plasticity of many of the known bacterial species, we can anticipate that such
49 methods can only partially identify genetic determinants of phenotypic variation.
50 To enable discovery of mechanisms related for instance to gene content,
51 alternative alignment-free methods have also been introduced^{1,4}. These methods
52 use k-mers, i.e. DNA words of length k, as generalized alternatives to SNPs as
53 putative explanations for observed differences in phenotype distributions. The
54 main advantage of k-mers is their ability to capture several different types of
55 variation present across a collection of genomes, including mutations,
56 recombinations, variable promoter architecture, differences in gene content as
57 well as capturing these variations in regions not present in all genomes.

58

59 The previous study using k-mers to overcome limitations of SNP-based
60 association used Monte-Carlo simulations of word gain and loss along an
61 inferred phylogeny to control for population structure¹, whereas SNP-based
62 studies have used clustering algorithms on a core alignment and stratified
63 association tests on the resulting groups of samples^{2,3}. The former does not scale
64 computationally to the hundreds of isolates required to find lower effect-size
65 associations, and the latter requires a core alignment, which lacks sensitivity and
66 difficult to produce when there is a large number of samples, or they are
67 particularly diverse.

68

69 Here we present a sequence element enrichment analysis (SEER), a method
70 computationally scalable to tens of thousands of genomes, implemented as a
71 stand-alone pipeline that uses either *de novo* assembled contigs or raw read data
72 as input. We apply SEER to both simulated and real data from a large and diverse
73 population, and show that it can accurately detect associations with antibiotic
74 resistance caused by both presence of a gene and by SNPs in coding regions.

75

76 **Results**

77 **Implementation**

78 SEER implements and combines three key insights which we discuss in turn: an
79 efficient scan of all possible k-mers with a distributed string mining algorithm,
80 an appropriate alignment-free correction for clonal population structure, and a
81 fast and fully robust association analysis of all counted k-mers.

82

83 K-mers allow simultaneous discovery of both short genetic variants and entire
84 genes associated with a phenotype. Longer k-mers provide higher specificity but
85 less sensitivity than shorter k-mers. Rather than arbitrarily selecting a length
86 prior to analysis or having to count k-mers at multiple lengths and combine the
87 results, we provide an efficient implementation that allows counting and testing
88 simultaneously at all k-mers at lengths over 9 bases long.

89

90 We offer three different methods to count k-mers in all samples in a study. For
91 very large studies, or for counting directly from reads rather than assemblies, we
92 provide an implementation of distributed string mining (DSM)^{6,7} which limits
93 maximum memory usage per core, but requires a large cluster to run. For data
94 sets up to around 5 000 sample assemblies we have implemented a single core
95 version fsm-lite (<https://github.com/nvalimak/fsm-lite>). For comparison with
96 older datasets, or where resources do not allow the storage of the entire k-mer
97 index in memory, DSK⁸ is used to count a single k-mer length in each sample
98 individually, the results of which are then combined.

99

100 To correct for the clonal population structure of bacterial populations, a distance
101 matrix is constructed from a random subsample of these k-mers, on which multi-
102 dimensional scaling is performed (Supplementary figure 1). Compared with
103 modelling SNP variation⁹, use of k-mers as variable sequence elements has been
104 previously shown to accurately estimate bacterial population structure. The
105 projections of each sample in three dimensions are used as covariates to control
106 for the clonal population structure. Simulations of bacterial genomes using a
107 known tree showed this method gave a higher resolution control than using only
108 population clustering (Supplementary figure 2). Before testing for association we
109 filter k-mers based on their frequency and unadjusted p-value to reduce false
110 positives from testing underpowered k-mers and reduce computational time.

111

112 Then, for each k-mer, a logistic curve is fitted to binary phenotype data, and a
113 linear model to continuous data, using a time efficient optimisation routine to
114 allow testing of all k-mers. Bacteria can be subject to extremely strong selection
115 pressures, producing common variants with very large effect sizes, such as
116 antibiotics inducing resistance-conferring variants. This can make the data
117 perfectly separable, and consequently the maximum likelihood estimate ceases
118 to exist for the logistic model. Firth regression¹⁰ has been used to obtain results
119 in these cases.

120

121 For the basal cut-off for significance we use $p < 0.05$, which in our testing we
122 conservatively Bonferroni corrected to the threshold 1×10^{-8} based on every
123 position in the *S. pneumoniae* genome having three possible mutations¹¹, and all
124 this variation being uncorrelated. This is a strict cut-off level that prevents a
125 large number of false-positives due to the extensive amount of k-mers being
126 tested, but does not over-penalise by correcting directly on the basis of the
127 number of k-mers counted. Simulations suggested a cut-off of 1.4×10^{-8} would be
128 appropriate, supporting this reasoning. Association effect size and p-value of the
129 MDS components can also be included in the output, to compare lineage and
130 variant effects on the phenotype variation.

131

132 K-mers reaching significance are filtered post-association and mapped onto both
133 a well-annotated reference sequence and the annotated draft assemblies to allow
134 discovery of variation in accessory genes not present in the reference strain. The
135 significant k-mers themselves can also be assembled into a longer consensus
136 sequence. Annotating variants by predicted function and effect (against a
137 reference sequence) in the resulting k-mers facilitates fine-mapping of SNPs and
138 small indels.

139

140 Meta-analysis of association studies increases sample size, which improves
141 power and reduces false-positive rates¹². To facilitate meta-analysis of k-mers
142 across studies, the output of SEER includes effect size, direction and standard
143 error, which can be used directly with existing software to meta-analyse all
144 overlapping k-mers.

145

146 SEER is implemented in C++, and available at <https://github.com/johnlees/seer>
147 as source code and a pre-compiled binary.

148 **Application to simulated data**

149 To test the power of SEER across different sample sizes, we simulated 3 069
150 *Streptococcus pneumoniae* genomes from the phylogeny observed in a Thai
151 refugee camp¹³ using parameters estimated from real data including
152 accumulation of SNPs, indels (Supplementary figure 3), gene loss and
153 recombination events. Using knowledge of the true alignments, we then
154 artificially associated an accessory gene with a phenotype over a range of odds-
155 ratios and evaluated power at different sample sizes (Fig. 1a). The expected
156 pattern for this power calculation is seen, with higher odds-ratio effects being
157 easier to detect. Currently detected associations in bacteria have had large effect
158 sizes (OR > 28 host-specificity¹, OR > 3 beta-lactam resistance²), and the required
159 sample sizes predicted here are consistent with these discoveries.

160

161 The large k-mer diversity, along with the population stratification of gene loss,
162 makes the simulated estimate of the sample size required to reach the stated
163 power clearly conservative. Convergent evolution along multiple branches of a

164 phylogeny for a real population reacting to selection pressures will reduce the
165 required sample size¹⁴.

166

167 We also used k-mers counted at constant lengths by DSK to perform the gene
168 presence/absence association (Fig. 1b). Counting all informative k-mers rather
169 than a range of pre-defined k-mer lengths gives greater power to detect
170 associations, with 80% power being reached at around 1 500 samples, compared
171 with 2 000 samples required by the pre-defined lengths. The slightly lower
172 power at low sample numbers is due to a stricter Bonferroni adjustment being
173 applied to the larger number of DSM k-mers over the DSK k-mers. This is exactly
174 the expected advantage from including shorter k-mers to increase sensitivity, but
175 as k-mers are correlated with each other due to evolving along the same
176 phylogeny, using the same Bonferroni correction for multiple testing does not
177 decrease specificity.

178

179 The strong linkage disequilibrium (LD) caused by the clonal reproduction of
180 bacterial populations means that non-causal k-mers may also appear to be
181 associated. This is well documented in human genetics; non-causal variants tag
182 the causal variant increasing discovery power, but make it more difficult to fine-
183 map the true link between genotype and phenotype¹⁵. In simulations it is difficult
184 to replicate the LD patterns observed in real populations, as recombination maps
185 for specific bacterial lineages are not yet known. To evaluate fine-mapping
186 power of a SNP we instead used the real sequence data and simulated
187 phenotypes based on changing the effect size of a known causal variant and
188 evaluating the physical distance of significant k-mers from the variant site.

189

190 Using DSM we counted 68M k-mers which we then tested for association. The
191 2 639 significant k-mers were placed into three categories if after mapping to a
192 reference genome they contained the causal variant I100L (10), were within the
193 same gene (74), or within 2.5kb in either direction (207). Figure 1c) shows the
194 resulting power when random subsamples of the population are taken. As
195 expected, power is higher when not specifying that the causal variant must be

196 hit, as there are many more k-mers which are in LD with the SNP than directly
197 overlapping it, thus increasing sensitivity.

198 **Confirmation of known resistance mechanisms in a large population of *S.***
199 ***pneumoniae***

200 SEER was applied to the sequenced genomes from the study described above,
201 using measured resistance to five different antibiotics as the phenotype:
202 chloramphenicol, erythromycin, β -lactams, tetracycline and trimethoprim.
203 Chloramphenicol resistance is conferred by the *cat* gene on the integrative
204 conjugative element (ICE) Tn5253 in the *S. pneumoniae* chromosome, and
205 similarly tetracycline resistance is conferred by the *tetM* gene which is also
206 carried on the ICE¹⁶. For both of these drug resistance phenotypes the ICE
207 contains 99% of the significant k-mers, and the causal genes rank highly within
208 the clusters (Table 1, Supplementary figure 4).

209

210 Resistance to erythromycin is also conferred by presence of a gene, but there are
211 multiple genes that can perform the same function (*ermB*, *mef*, *mel*)¹⁷. In the
212 population studied, this phenotype was strongly associated with two large
213 lineages (Supplementary figure 5), making the task of disentangling association
214 with a lineage versus a specific locus more difficult. Significant k-mers are found
215 in the mega and omega cassettes, which carry the *mel/mef* and *ermB* resistance
216 elements respectively. Some k-mers do not map to the reference, as they are due
217 to lineage specific associations with genetic elements not found in the reference
218 strain. This highlights both the need to map to a close reference or draft
219 assembly to interpret hits, as well as the use of functional follow-up to validate
220 potential hits from SEER.

221

222 Multiple mechanisms of resistance to β -lactams are possible². Here, we consider
223 just the most important (i.e. highest effect size) mutations, which are SNPs in the
224 penicillin binding proteins *pbp2x*, *pbp2b* and *pbp1a*. In this case looking at
225 highest coverage annotations finds these genes, but is not sufficient as so many
226 k-mers are significant – either due to other mechanisms of resistance, physical
227 linkage with causal variants or co-selection for resistance conferring mutations.
228 Instead, looking at the k-mers with the most significant p-values gives the top

229 four hit loci as *pbp2b* ($p=10^{-132}$), *pbp2x* ($p=10^{-96}$), putative RNA pseudouridylate
230 synthase UniParc B8ZPU5 ($p=10^{-92}$) and *pbp1a* ($p=10^{-89}$). The non-*pbp* hit is a
231 homologue of a gene in linkage disequilibrium with *pbp2b*, which would suggest
232 mismapping rather than causation of resistance.

233

234 Trimethoprim resistance in *S. pneumoniae* is conferred by the SNP I100L in the
235 *folA/dyr* gene¹⁸. The *dpr* and *dpr* genes, which are adjacent in the genome, have
236 the highest coverage of significant k-mers (Fig. 2). Following our fine-mapping
237 procedure, we call four high-confidence SNPs that are predicted to be more likely
238 to affect protein function than synonymous SNPs. One is the causal SNP, and the
239 others appear to be hitchhikers in LD with I100L. By evaluating whether sites are
240 conserved across the protein family¹⁹, the known causal SNP is ranked as the
241 highest variant, showing that in this case fine-mapping is possible using the
242 output from SEER.

243

244 We then compared the results from SEER with the results from two existing
245 methods (as described in online methods). The first method uses mapping of
246 SNPs against a reference, followed by applying the Cochran–Mantel–Haenszel
247 test at every variable site². The second uses *dsk*⁸ to count k-mers of length 31,
248 and a highly robust correction for population structure which scales to around
249 100 genomes¹.

250

251 The results are shown in supplementary table 1. Both SEER and association of a
252 core mapping of SNPs identify resistances caused by presence of a gene, when it
253 is present in the reference used for mapping. Both produce their most significant
254 p-values in the causal element, though SEER appears to have a lower false-
255 positive rate. However, as demonstrated by chloramphenicol resistance, if not
256 enough SNP calls are made in the causal gene this hinders fine-mapping. SNP-
257 mediated resistance showed the same pattern since many other SNPs were
258 ranked above the causal variant. In the case of β -lactam resistance both methods
259 seem to perform equally well, likely due to the higher rate of recombination and
260 the creation of mosaic *pbp* genes.

261

262 Additionally, as for erythromycin resistance, when an element is not present in
263 the reference SNPs have been called against it is not detectable in SNP-based
264 association analysis. In such cases multiple mappings against other reference
265 genomes would have to be made, which is a tedious and computationally costly
266 procedure. Alternatively a draft assembly with the phenotype from the study
267 could be picked as a second reference to map to, however this may be lower
268 quality than those in public databases picked by genetic content rather than
269 phenotype, and would not necessarily be able to detect multiple genetic
270 mechanisms (as in the case of erythromycin resistance, no single sequenced
271 genome contains all known resistance mechanisms).

272

273 Since the k-mer results from SEER are reference-free, these issues are avoided as
274 just the significant k-mers can quickly be mapped to all available references.
275 Alternatively, the significant k-mers can be mapped to all draft assemblies in the
276 study, at least one of which is guaranteed to contain the k-mer, to check if any
277 annotations are overlapped.

278

279 For the small sample, 31mer approach significance was not reached for
280 chloramphenicol, tetracycline or trimethoprim as the effect size of any k-mer is
281 too small to be detected in the number of samples accessible by the method.
282 Erythromycin had 19 307 hits, and β -lactams 419 hits, at between 1-2% MAF
283 which are all false positives that would likely have been excluded by a fully
284 robust population structure correction method.

285 **Discussion**

286 SEER is a reference-independent, scalable pipeline capable of finding bacterial
287 sequence elements associated with a range of phenotypes while controlling for
288 clonal population structure. The sequence elements can be interpreted in terms
289 of protein function using sequence databases, and we have shown that even
290 single causal variants can be fine-mapped using the SEER output.

291

292 Our use of all informative k-mers together with robust regression methods, and
293 the ability to analyse very large sample sizes show improved sensitivity over

294 existing methods. This provides a generic approach capable of analysing the
295 rapidly increasing number of bacterial whole genome sequences linked with a
296 range of different phenotypes. The output can readily be used in a meta-analysis
297 of sequence elements to facilitate the combination of new studies with published
298 data, increasing both discovery power and confirming the significance of results.
299 As with all association methods, our approach is limited by the amount of
300 recombination and convergent evolution that occurs in the observed population,
301 since the discovery of causal sequence elements is principally constrained by the
302 extent of linkage disequilibrium. However, by introducing improved
303 computational scalability and statistical sensitivity SEER significantly pushes the
304 existing boundaries for answering important biologically and medically relevant
305 questions.

306 **Online methods**

307 **Counting informative k-mers in samples**

308 Over all N samples, all k-mers over 9 bases long that occur in more than one
309 sample are counted. All non-informative k-mers are omitted from the output; a
310 k-mer X is not informative if any one base extension to the left (aX) or right (Xa)
311 has exactly the same frequency support vector as X . The frequency support
312 vector has N entries, each being the number of occurrences of k-mer X in that
313 sample. Further filtering conditions are explained in the sections below.

314

315 Distributed string mining (DSM)^{6,7} parallelises to as much as one sample per
316 core, and either 16 or 64 master server processes. DSM includes an optional
317 entropy-filtering setting that filters the output k-mers based on both number of
318 samples present and frequency distribution. On our 3 069 simulated genomes
319 this took 2 hrs 38 min on 16 cores, and used 1Gb RAM. The distributed approach
320 is applicable up to terabytes of short-read data⁷, but requires a cluster
321 environment to run. As an easy-to-use alternative, we propose a single core
322 version of DSM that is applicable for gigabyte-scale data. We implemented the
323 single core version based on a succinct data structure library²⁵ to produce the
324 same output as DSM. On 675 *S. pyogenes* genomes this took 3hrs 44min and used
325 22.3Gb RAM.

326

327 To count single k-mer lengths, an associative array was used to combine the
328 results from DSK in memory. We concatenated results from k-mer lengths of 21,
329 31 and 41, as in previous studies¹. This can scale to large genome numbers by
330 instead using external sorting to avoid storing the entire array in memory.

331 **Filtering k-mers**

332 K-mers are filtered if either they appear in <1% or >99% of samples, or are over
333 100 bases long. We also test if the p-value of association in a simple χ^2 test (1
334 d.f.) is less than 10^{-5} , as in simulations this was true for all true positives. In the
335 case of a continuous phenotype a Welch two-sample t-test is used instead.

336 **Covariates to control for population structure**

337 A random sample of between 0.1% and 1% of k-mers appearing in between 5-
338 95% of isolates is taken. We then construct a pairwise distance matrix **D**, with
339 each element being equal to a sum over all m sampled k-mers:

$$d_{ij} = \sum_m \|k_{im} - k_{jm}\|$$

340 where k_{im} is 1 if the m th sampled k-mer is present in sample i , and 0 otherwise.

341

342 Metric multi-dimensional scaling is then performed, projecting these distances
343 into three dimensions. The normalised eigenvectors of each dimension are used
344 as covariates in the regression model. The number of dimensions used is a user-
345 adjustable parameter, and can be evaluated by the goodness-of-fit and the
346 magnitude of the eigenvalues. In species tree with two lineages and 96 isolates
347 one dimension was sufficient as a population control, whereas for the larger
348 collection of 3069 isolates 10-15 dimensions were needed to give tight control
349 (Supplementary figure 6). Over all our studies, generally three dimensions
350 appeared a good trade-off between sensitivity and specificity.

351 **Logistic and linear regression**

352 For samples with binary outcome vector \mathbf{y} , for each k-mer a logistic model is
353 fitted:

$$\log\left(\frac{\mathbf{y}}{I - \mathbf{y}}\right) = \mathbf{X}\boldsymbol{\beta}$$

354 where absence and presence for each k-mer coded as 0 and 1 respectively in
355 column 2 of the design matrix \mathbf{X} (column 1 is a vector of ones, giving an intercept
356 term). Subsequent columns j of \mathbf{X} contain the eigenvectors of the MDS projection,
357 user-supplied categorical covariates (dummy encoded), and quantitative
358 covariates (normalised). The BFGS algorithm is used to maximise the log
359 likelihood L in terms of the gradient vector $\boldsymbol{\beta}$ (using an analytic expression for
360 $d(\log L)/d\boldsymbol{\beta}$):

$$\log L \propto \sum_i y_i \cdot \log(\text{sig}(\mathbf{X}\boldsymbol{\beta})_i) + (1 - y_i) \cdot \log(\text{sig}(1 - \mathbf{X}\boldsymbol{\beta})_i)$$

361 where sig is the sigmoid function. If this fails to converge, n Newton-Raphson
362 iterations are applied to $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}_{n+1} = \boldsymbol{\beta}_n + [-L''(\boldsymbol{\beta}_n)]^{-1} \cdot L'(\boldsymbol{\beta}_n)$$

363 from a starting point using the mean phenotype as the intercept, and the root-
364 mean squared beta from a test of k-mers passing filtering

$$\beta_{0,0} = \frac{\sum y_i}{n}$$
$$\beta_{0,j>0} = 0.1$$

365 which is slower, but has a higher success rate. If this fails to converge due to the
366 observed points being separable in the high dimensional space, or the standard
367 error of the slope is greater than 3 (which empirically indicated almost separable
368 data, with no counts in one element of the contingency table), Firth logistic
369 regression¹⁰ is then applied. This adds an adjustment to log L :

$$\log L(\boldsymbol{\beta})^* = \log L(\boldsymbol{\beta}) + \frac{1}{2} \cdot \log \left| \frac{d^2 L}{d\boldsymbol{\beta}^2}(\boldsymbol{\beta}) \right|$$

370 using which Newton-Raphson iterations are applied as above.

371

372 In the case of a continuous phenotype a linear model is fitted:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$$

373 The squared distance $U(\boldsymbol{\beta})$

$$U(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

374 is minimised using the BFGS algorithm. If this fails to converge then the analytic
375 solution is obtained by orthogonal decomposition:

$$\mathbf{X} = \mathbf{Q}\mathbf{R}$$

376 then back-solving for $\boldsymbol{\beta}$ in:

$$R\beta = Q^T y$$

377

378 In both cases the standard error on β_1 is calculated by inverting the Fisher
379 information matrix $d^2L/d\beta^2$ (inversions are performed by Cholesky
380 decomposition, or if this fails due to the matrix being almost singular the Moore-
381 Penrose pseudoinverse is taken) to obtain the variance-covariance matrix. The
382 Wald statistic is calculated with the null hypothesis of no association ($\beta_1 = 0$):

$$W = \frac{\beta_1}{SE(\beta_1)}$$

383 which is the test statistic of a χ^2 distribution with 1 d.f. This is equivalent to the
384 positive tail of a standard normal distribution, the integral of which gives the p-
385 value. To calculate an empirical significance testing cut-off for the p-value under
386 multiple correlated tests, we observed the distribution of p-values from 100
387 random permutations of phenotype. Setting the family-wise error rate (FWER) at
388 0.05 gave a cut-off of 1.4×10^{-8} .

389 SEER implementation

390 SEER is implemented in C++ using the armadillo linear algebra library²⁶, and dlib
391 optimisation library²⁷. On a simulation of 3 069 diverse 0.4Mb genomes, 143M k-
392 mers were counted by DSM and 25M 31-mers by DSK. On the largest DSM set,
393 using 16 cores and subsampling 300 000 k-mers (0.2% of the total), calculating
394 population covariates took 6hr 42min and 8.33GB RAM. This step is $O(N^2M)$
395 where N is number of samples and M is number of k-mers, but can be
396 parallelised across up to N^2 cores.

397

398 Processing all 143M informative k-mers as described took 69min 44s and 23MB
399 RAM on 16 cores. This step is $O(M)$ and can be parallelised across up to M cores.

400

401 On the real dataset of full length genomes the 68M informative k-mers counted
402 was less than the simulated dataset above, as the parameters of the simulation
403 created particularly diverse final genomes.

404 **Interpreting significant k-mers**

405 K-mers reaching the threshold for significance are then post-association filtered
406 requiring $\beta_1 > 0$ as a negative effect size does not make biological sense.

407 Remaining k-mers are searched for by exact match in their *de novo* assemblies,
408 and annotations of features examined for overlap of function. BLAT²⁸ is also used
409 with a step size of 2 and minimum match size of 15 to find inexact but close
410 matches to a well annotated reference sequence.

411

412 To better search for gene clusters associated with phenotype, these k-mers are
413 assembled using Velvet²⁹ choosing a smaller sub-k-mer size which maximises
414 longest contig length of the final assembly. K-mers which are then substrings of
415 others significant k-mers are removed.

416 **Mapping of a single SNP**

417 Using the BLAT mapping of significant k-mers to a reference sequence, SNPs are
418 called using bcftools³⁰. Quality scores for a read are set to be identical, and are
419 set as the Phred-scaled Holm-adjusted p-values from association. High quality
420 (QUAL > 100) SNPs are then annotated for function using SnpEff³¹, and the effect
421 of missense SNPs on protein function is ranked using SIFT¹⁹.

422 **Comparison to existing methods**

423 We compare to two existing methods. The first uses a core-genome SNP mapping
424 along with population clusters defined from the same alignment to perform a
425 Cochran-Mantel-Haenszel test at every called variant site². The second uses a
426 fixed k-mer length of 31 as counted by dsk⁸, with a Monte Carlo phylogeny-based
427 population control¹. As the second method is not scalable to this population size
428 we used our population control as calculated from all genomes in the population,
429 and a subsample of 100 samples to calculate association statistics, which is
430 roughly the number computationally accessible by this method. In both cases,
431 the same Bonferroni correction is used as for SEER.

432 **Simulating bacterial populations**

433 A random subset of 450 genes from the *Streptococcus pneumoniae* ATCC
434 700669¹⁶ strain were used as the starting genome for ALF³². ALF simulated 3069

435 final genomes along the phylogeny observed in a Thai refugee camp¹³. An
436 alignment between *S. pneumoniae* strains R6, 19F and *Streptococcus mitis* B6
437 using Progressive Cactus was used to estimate rates in the GTR matrix and the
438 size distribution of insertions and deletions (INDELs – Supplementary figure 3).
439 Previous estimates for the relative rate of SNPs to INDELs³³ and the rate of
440 horizontal gene transfer and loss¹³ were used.
441 pIRS³⁴ was used to simulate error-prone reads from genomes at the tips of the
442 tree, which were then assembled by Velvet²⁹. DSM was used to count k-mers
443 from these *de novo* assemblies.

444
445 To test the similarity of the population control to existing methods, 96 full
446 *Streptococcus pneumoniae* ATCC 700669 genomes were evolved with ALF.
447 Intergenic regions were also evolved using Dawg³⁵ at a previously determined
448 rate³⁶. These were combined, and assemblies generated and k-mers counted as
449 above. A distance matrix was created from 1% of the k-mers as described above,
450 and a neighbour-joining tree produced from this.

451
452 The resulting tree was ranked against the true tree by counting one for each pair
453 of isolates in each BAPS³⁷ cluster which had an isolate not in the same BAPS
454 cluster as a descendent of their MRCA.

455 **Simulating phenotype based on genotype and odds-ratio**

456 Ratio of cases to controls in the population (S_R) was set at 50% to represent
457 antibiotic resistance, and a single variant (gene presence/absence or a SNP) was
458 designated as causal. Minor allele frequency (MAF) in the population is set from
459 the simulation, and odds-ratio (OR) can be varied. The number of disease cases
460 D_E is then the solution to a quadratic equation³⁸, which is related to probability of
461 a sample being a case by:

$$p_{\text{case|exposed}} = \frac{D_E}{\text{MAF}}$$
$$p_{\text{case|not exposed}} = \frac{\frac{S_R}{S_R + 1} - D_E}{1 - \text{MAF}}$$

462 The population was then randomly subsampled 100 times, with case and control
463 status assigned for each run using these formulae. Power was defined by the

464 proportion of runs that had at least one k-mer in the gene associated with
465 phenotype reaching significance.

466 **References**

467

- 468 1. Sheppard, S. K. *et al.* Genome-wide association study identifies vitamin B5
469 biosynthesis as a host specificity factor in *Campylobacter*. *Proc. Natl. Acad.*
470 *Sci.* **110**, 11923–11927 (2013).
471
- 472 2. Chewapreecha, C. *et al.* Comprehensive Identification of Single Nucleotide
473 Polymorphisms Associated with Beta-lactam Resistance within
474 Pneumococcal Mosaic Genes. *PLoS Genet.* **10**, e1004547 (2014).
475
- 476 3. Laabei, M. *et al.* Predicting the virulence of MRSA from its genome
477 sequence. *Genome Res.* **24**, 839–849 (2014).
478
- 479 4. Weinert, L. a. *et al.* Genomic signatures of human and animal disease in the
480 zoonotic pathogen *Streptococcus suis*. *Nat. Commun.* **6**, 6740 (2015).
481
- 482 5. Chen, P. E. & Shapiro, B. J. The advent of genome-wide association studies
483 for bacteria. (2015).
484
- 485 6. Välimäki, N. & Puglisi, S. in *Algorithms Bioinforma. SE - 35* (Raphael, B. &
486 Tang, J.) **7534**, 441–452 (Springer Berlin Heidelberg, 2012).
487
- 488 7. Seth, S., Välimäki, N., Kaski, S. & Honkela, A. Exploration and retrieval of
489 whole-metagenome sequencing samples. *Bioinformatics* **30**, 16 (2014).
490
- 491 8. Rizk, G., Lavenier, D. & Chikhi, R. DSK: K-mer counting with very low
492 memory usage. *Bioinformatics* **29**, 652–653 (2013).
493
- 494 9. Tasoulis, S. *et al.* Random projection based clustering for population
495 genomics. in *2014 IEEE Int. Conf. Big Data (Big Data)* 675–682 (2014).
496 doi:10.1109/BigData.2014.7004291
497
- 498 10. Heinze, G. & Schemper, M. A solution to the problem of separation in
499 logistic regression. *Stat. Med.* **21**, 2409–2419 (2002).
500
- 501 11. Ford, C. B. *et al.* Mycobacterium tuberculosis mutation rate estimates from
502 different lineages predict substantial differences in the emergence of drug-

- 503 resistant tuberculosis. *Nat. Genet.* **45**, 784–790 (2013).
504
- 505 12. Evangelou, E. & Ioannidis, J. P. A. Meta-analysis methods for genome-wide
506 association studies and beyond. *Nat. Rev. Genet.* **14**, 379–389 (2013).
507
- 508 13. Chewapreecha, C. *et al.* Dense genomic sampling identifies highways of
509 pneumococcal recombination. *Nat. Genet.* **46**, 305–9 (2014).
510
- 511 14. Farhat, M. R. *et al.* Genomic analysis identifies targets of convergent
512 positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat.*
513 *Genet.* **45**, 1183–9 (2013).
514
- 515 15. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Hum.*
516 *Mol. Genet.* **24**, R111–R119 (2015).
517
- 518 16. Croucher, N. J. *et al.* Role of conjugative elements in the evolution of the
519 multidrug-resistant pandemic clone *Streptococcus pneumoniae* Spain23F
520 ST81. *J. Bacteriol.* **191**, 1480–1489 (2009).
521
- 522 17. Croucher, N. J. *et al.* Rapid pneumococcal evolution in response to clinical
523 interventions. *Science* **331**, 430–4 (2011).
524
- 525 18. Maskell, J. P., Sefton, a. M. & Hall, L. M. C. Multiple mutations modulate the
526 function of dihydrofolate reductase in trimethoprim-resistant
527 *Streptococcus pneumoniae*. *Antimicrob. Agents Chemother.* **45**, 1104–1108
528 (2001).
529
- 530 19. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect
531 protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
532
- 533 20. Roberts, A. P. & Mullany, P. A modular master on the move: the Tn916
534 family of mobile genetic elements. *Trends Microbiol.* **17**, 251–258 (2009).
535
- 536 21. Dubnau, D. DNA Uptake in Bacteria. *Annu. Rev. Microbiol.* **53**, 217–244
537 (1999).
538
- 539 22. Lefébure, T. & Stanhope, M. J. Evolution of the core and pan-genome of
540 *Streptococcus*: positive selection, recombination, and genome
541 composition. *Genome Biol.* **8**, R71 (2007).
542
- 543 23. Raeder, R. & Boyle, M. D. Association between expression of

- 544 immunoglobulin G-binding proteins by group A streptococci and virulence
545 in a mouse skin infection model. *Infect. Immun.* **61**, 1378–1384 (1993).
546
- 547 24. Raeder, R. & Boyle, M. D. Analysis of immunoglobulin G-binding-protein
548 expression by invasive isolates of *Streptococcus pyogenes*. *Clin. Diagn. Lab.*
549 *Immunol.* **2**, 484–486 (1995).
550
- 551 25. Gog, S., Beller, T., Moffat, A. & Petri, M. in *Exp. Algorithms SE - 28*
552 (Gudmundsson, J. & Katajainen, J.) **8504**, 326–337 (Springer International
553 Publishing, 2014).
554
- 555 26. Sanderson, C. Armadillo: An Open Source C++ Linear Algebra Library for
556 Fast Prototyping and Computationally Intensive Experiments. in *NICTA*
557 **NICTA**, 1–16 (2010).
558
- 559 27. King, D. E. Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* **10**,
560 1755–1758 (2009).
561
- 562 28. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–
563 664 (2002).
564
- 565 29. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read
566 assembly using de Bruijn graphs. *Genome Res.* **18**, 821–9 (2008).
567
- 568 30. Li, H. A statistical framework for SNP calling, mutation discovery,
569 association mapping and population genetical parameter estimation from
570 sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
571
- 572 31. Cingolani, P. *et al.* A program for annotating and predicting the effects of
573 single nucleotide polymorphisms, SnpEff: SNPs in the genome of
574 *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. **6**, 1–13
575 (2012).
576
- 577 32. Dalquen, D. a, Anisimova, M., Gonnet, G. H. & Dessimoz, C. ALF--a
578 simulation framework for genome evolution. *Mol. Biol. Evol.* **29**, 1115–23
579 (2012).
580
- 581 33. Chen, J. Q. *et al.* Variation in the ratio of nucleotide substitution and indel
582 rates across genomes in mammals and bacteria. *Mol. Biol. Evol.* **26**, 1523–
583 1531 (2009).
584

- 585 34. Hu, X. *et al.* pIRS: Profile-based Illumina pair-end reads simulator.
586 *Bioinformatics* **28**, 1533–1535 (2012).
587
- 588 35. Cartwright, R. a. DNA assembly with gaps (Dawg): Simulating sequence
589 evolution. *Bioinformatics* **21**, 31–38 (2005).
590
- 591 36. Kosiol, C., Holmes, I. & Goldman, N. An empirical codon model for protein
592 sequence evolution. *Mol. Biol. Evol.* **24**, 1464–1479 (2007).
593
- 594 37. Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M. & Corander, J.
595 Hierarchical and spatially explicit clustering of DNA sequences with BAPS
596 software. *Mol. Biol. Evol.* **30**, 1224–8 (2013).
597
- 598 38. Newman, S. C. in *Biostat. Methods Epidemiol.* 329–330 (John Wiley & Sons,
599 Inc., 2003). doi:10.1002/0471272612.app4
600

601 **Acknowledgements**

602 We would like to thank James Hadfield for his help in integrating SEER's output
603 into the bacterial genome visualisation tool JSCandy, and Jeff Barrett and his
604 group for helpful discussions on the relation of association studies in human
605 genetics to prokaryotic genetics.

606 This work was supported by Wellcome Trust grant 098051, MRC grant 1365620,
607 ERC grant 239784, Academy of Finland grant 287665 and COIN Centre of
608 Excellence.

609 **Competing Interests**

610 The authors declare no competing interests.

611 **Author Contributions**

612 JAL – Designed method, performed analysis, wrote manuscript.

613 MV – Designed method, performed analysis, wrote manuscript.

614 NV – Participated in method design, edited manuscript.

615 SRH – Helped with interpretation data

616 CC – Prepared genetic and metadata from Maela isolates.

617 NJC – Helped with interpretation of antibiotic resistance elements, edited

618 manuscript.

619 PM – Participated in method design, edited manuscript.
620 AH – Participated in method design, edited manuscript.
621 JP – Advised on microbiological interpretation, edited manuscript.
622 SDB – Advised on microbiological interpretation, edited manuscript.
623 JC - Designed method, performed analysis, wrote manuscript.

624 Data Access

625 SEER is available at <https://github.com/johnlees/seer>, DSM at
626 <https://github.com/HIITMetagenomics/dsm-framework> and fsm-lite at
627 <https://github.com/nvalimak/fsm-lite>.
628 Scripts used to perform the simulations are available at
629 <https://github.com/johnlees/bioinformatics>

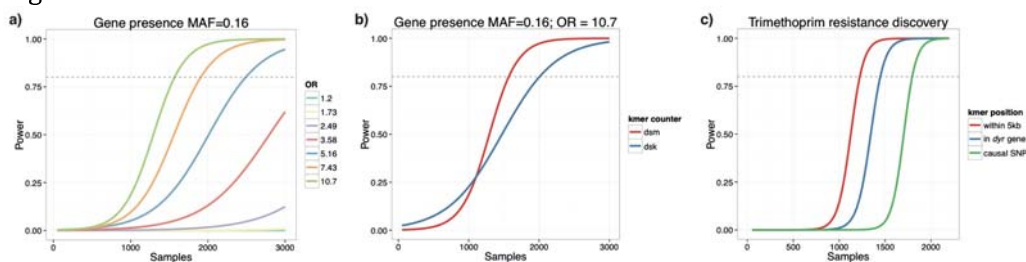
630 Figure Captions

631 Fig. 1: Using simulations and subsamples of the population as described in the
632 online methods, power for a) detecting gene presence/absence at different odds-
633 ratios b) using all informative k-mers versus a single length c) detecting k-mers
634 near, in the correct gene, or containing the causal variant for trimethoprim
635 resistance. All curves are logistic fits to the mean power over 100 subsamples.
636

637 Fig. 2: Fine mapping trimethoprim resistance. The locus pictured contains 72
638 significant k-mers, the most of any gene cluster. Coverage over the locus is
639 pictured at the bottom of the figure. Shown above the genes are high quality
640 missense SNPs, plotted using their p-value for affecting protein function as
641 predicted by SIFT.

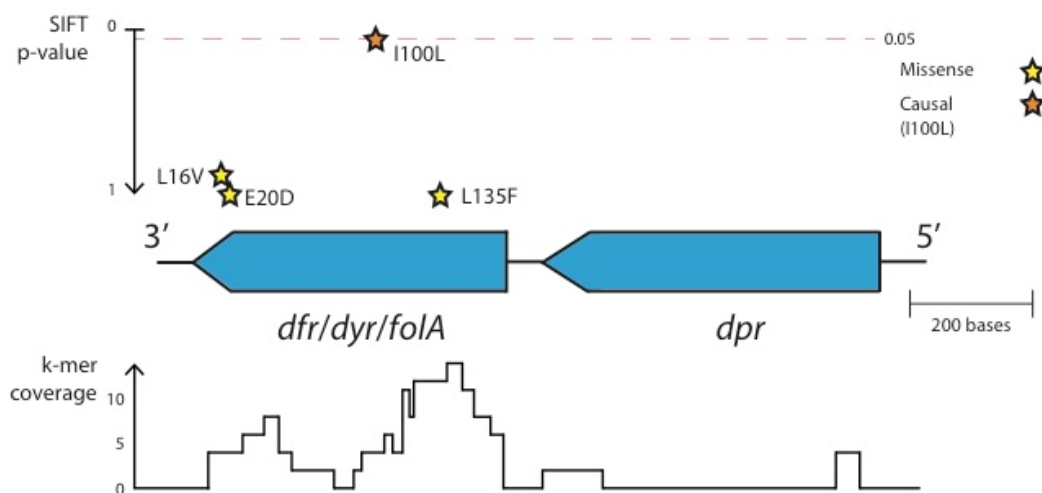
642 Figures

643 Fig. 1



644
645

Fig. 2



646

647 Tables

Antibiotic	Resistant samples	Number of significant k-mers			
		Total	Mapped to reference	Highest coverage annotation	Causal element
Chloramphenicol	204 (7%)	1 526	1 526	1 508 – ICE 288 – ORF (UniParc B8ZK82) 206 – <i>rep</i> 166 – <i>cat</i>	166 – <i>cat</i>
Erythromycin	803 (26%)	1 154	112	10 – permease (UniParc B8ZKV5) 8 – <i>prfC</i> 6 – <i>gatA</i> 4 – ICE	4 – mega element 2 – <i>mef</i> 2 – omega element
β -lactams	1 563 (51%)	23 876	17 453	381 – ICE 145 – prophage MM1 50 – SPN23F15110 (UniParc B8ZLE7) 49 – ICE <i>orf16</i>	47 – <i>pbp2x</i> 20 – <i>pbp2b</i> 8 – <i>pbp1a</i>
Tetracycline	1 958 (64%)	962	962	962 – ICE 136 – ICE <i>orf16</i> 121 – ICE <i>orf15</i> 96 – <i>tetM</i>	96 – <i>tetM</i>
Trimethoprim	2 553 (83%)	2 639	210	96 – <i>tetM</i> 21 – <i>dyr</i>	21 – <i>dyr</i>

648

649 Table 1: Results from SEER for antibiotic resistance binary outcome on a
 650 population of 3069 *S. pneumoniae*. Significant k-mers are first interpreted by
 651 mapping to the ATCC 700669 reference genome. Up to the first four highest
 652 covered annotations are shown, and if the known mechanism is amongst these it
 653 is highlighted in orange. The ICE is the top hit in three analyses, as it carries
 654 multiple drug-resistance elements and is commonly found in multi-drug
 655 resistant strains¹⁶. The distribution of phenotype across the phylogeny is shown
 656 in Supplementary figure 5.

657

658 **Supplementary data**

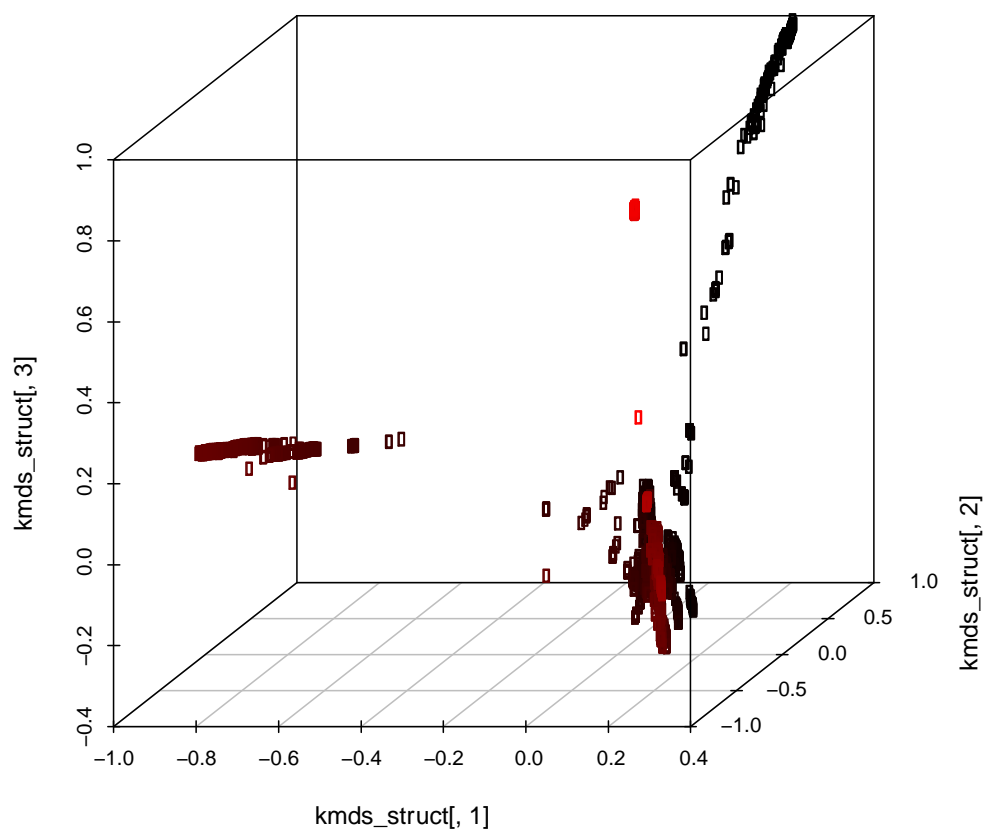
659

660 **Supplementary table 1:** Comparison of SEER with results from existing
 661 methods in finding genetic associations with antibiotic resistance in the
 662 Chewapreecha *et. al.* study of 3069 Thai carriage *S. pneumoniae* samples. For
 663 each of the five antibiotics, the true causal variant is listed, as are the number of
 664 hits passing the significance threshold for each method (plink and dsk) and the
 665 number which map to the correct region.
 666

Antibiotic	Causal variant	Significant sites		Near correct site		Notes
		plink	dsk	plink		
Tetracycline	ICE, <i>tetM</i>	8 029	0	<i>tetM</i> – 124	ICE – 2240	
Chloramphenicol	ICE, <i>cat</i>	5 310	0	<i>cat</i> – 0	ICE – 1137	
β-lactams	<i>pbp2x</i> , <i>pbp1a</i> , <i>pbp2b</i>	858	0	<i>pbp2x</i> – 210	<i>pbp1a</i> – 113	<i>pbp2b</i> – 81
Trimethoprim	<i>dyr</i> (I100L)	4 009	0	<i>dyr</i> – 47	<i>dpr</i> – 53	Causal SNP ranked 22nd
Erythromycin	<i>ermB</i> , <i>mef</i> , <i>mel</i> , <i>mefA</i>	8 469	0	None		Element not present in reference

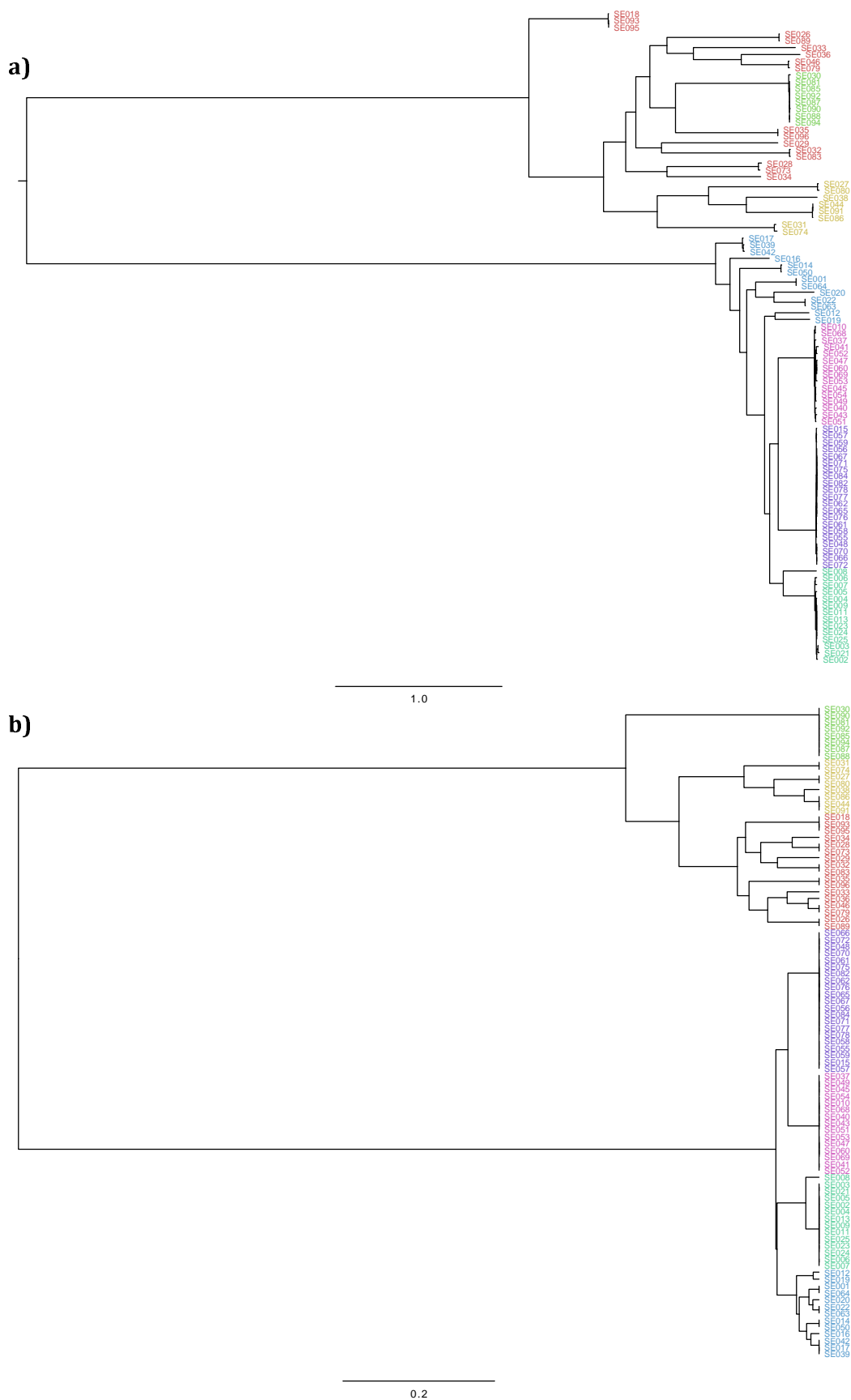
667

668



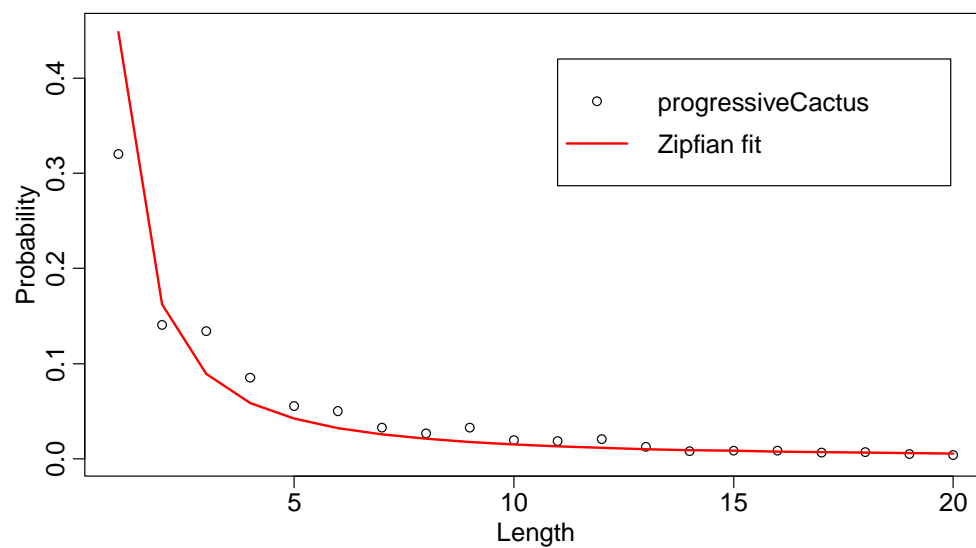
669

670 **Supplementary figure 1:** Plot of the k-mer distances projected into three
671 dimensions by MDS for the Chewapreecha *et. al.* study of 3069 Thai carriage *S.*
672 *pneumoniae* samples. Shade from black to red is by y-coordinate (2nd MDS
673 component).
674



675 **Supplementary figure 2:** a) Tree used for Monte Carlo simulations of 96 *S.*
676 *pneumoniae* genomes. b) UPGMA tree from k-mer distance matrix produced from

677 simulated reads. Colours are hierBAPS clusters.
678
679

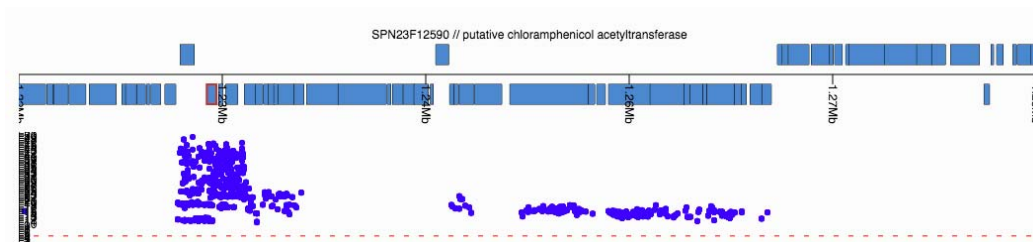


680

681 **Supplementary figure 3:** Estimated size distribution for INDELs, as estimated
682 from a Progressive Cactus alignment of three members of the *Streptococcus*
683 genus. A power law $p=L^k$ (Zipfian function; p is probability, L is INDEL length, k is
684 a free parameter) is fit to the data, the parameter k is used in the simulations.

685

686



687

688

689 **Supplementary figure 4:** JScandy view of ATCC 700669 reference genome (blue

690 blocks at top genes on forward and reverse strands) and Manhattan plot of start

691 positions of the 1 508 of 1 526 k-mers significantly associated with

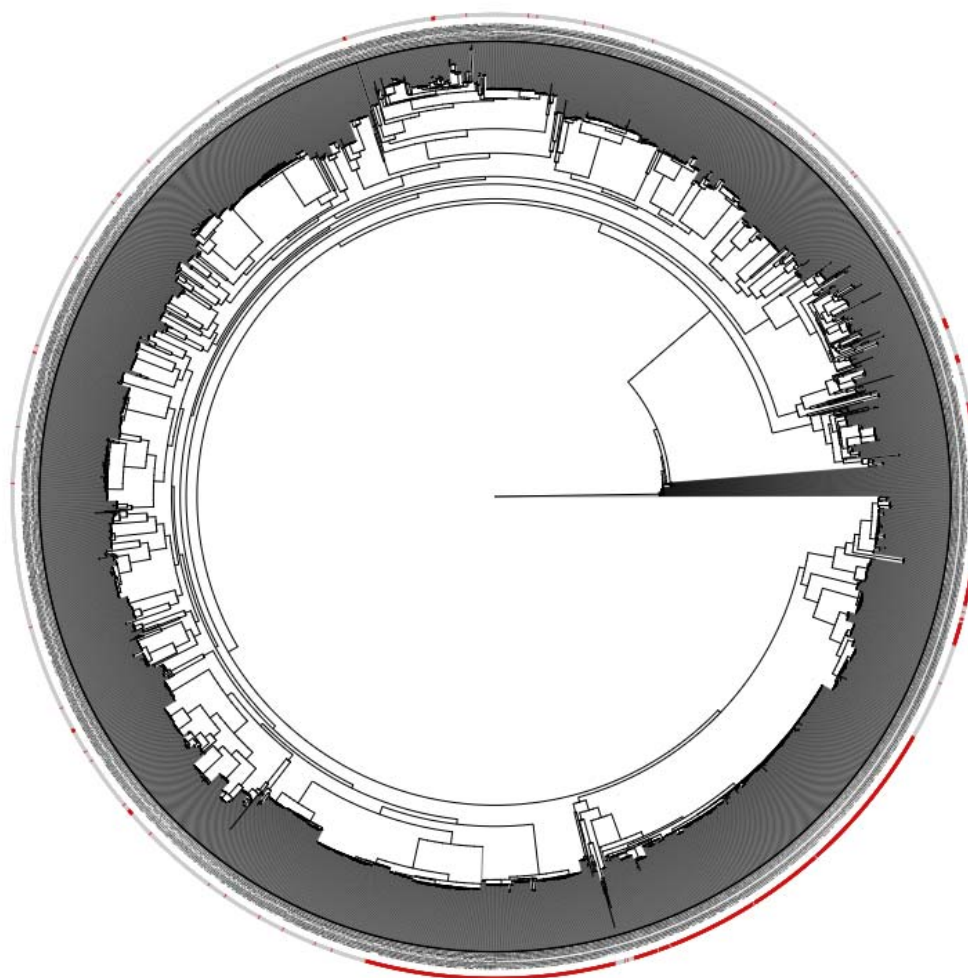
692 chloramphenicol resistance which map to the integrative conjugative element

693 (ICE) Tn5253. The hits are all in within the ICE, and the most significant hits

694 cluster around the *cat* gene (which is outlined in red).

695

696



697

698

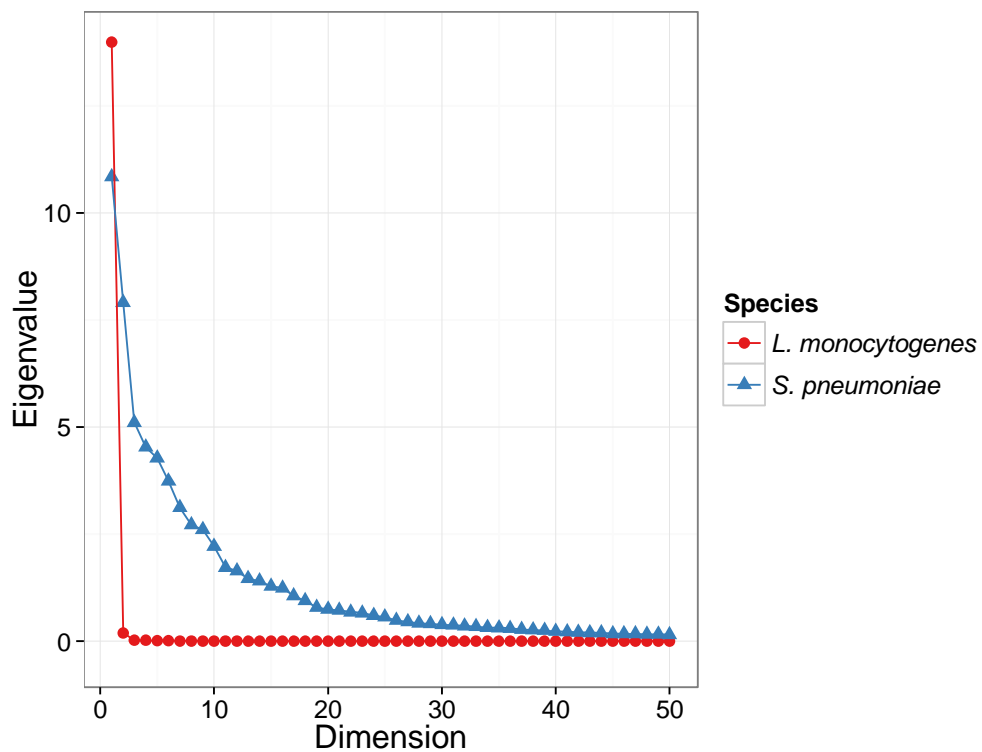
699

700

701

702

Supplementary figure 5: Neighbour joining tree from Chewapreecha *et. al.* study of 3069 Thai carriage *S. pneumoniae* samples, from a SNP alignment produced by mapping to the ATCC 700669 reference strain. Outer ring: red if resistant to Erythromycin, grey if sensitive.



703

704

Supplementary figure 6: Scree plot for the first fifty dimensions of the 96

705

Listeria monocytogenes isolates (Supplementary figure 2) in red, 3 069

706

Streptococcus pneumoniae isolates (Supplementary figure 5) in blue.

707