

Deliverable

05.2 Tools for rapid sequence-based detection of strain specific clusters in time, place and host for the main emerging pathogen classes

Version: 1

Due: Month 30

Completed: Month 51

Authors: Ole Lund (DTU), Judit Szarvas (DTU)

Contributing Partners: EMC, DTU



This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 643476.



Contents

Deliverable Description	2
Background	3
Evergreen platform.....	3
The solution	3



Deliverable Description

In Workpackage 5, 'From comparable data to actionable information: Additional tools for detection of and response to (re-) emerging infections', the analytical workflows developed in this workpackage aim to answer essential basic questions such as:

- Which pathogen is causing the emergence or outbreak, and what are its characteristics (e.g. virulence, transmission routes, toxins, antibiotic resistance)?
- Where and what is the original source of the pathogen?
- How do people get infected, and do they pass it on to others?
- What can we do to stop the emergence/outbreak or limit the impact?

While some of these are essentially the same questions as the ones addressed in Workpackage 4, the pathogens targeted here are not or only partially covered by routine diagnostics and surveillance programs or entirely unknown. Therefore, while Workpackages 3 and 4 focus on potential added value of NGS/WGS based approaches in the context of existing diagnostics and surveillance with current or improved sampling strategies designed for these surveillance systems (ECDC, EFSA, WHO), Workpackage 5 seeks to take a step further by working on novel applications to enhance emerging disease detection and investigation.

The submission of this deliverable has been delayed due to a lack of clarity as to the assignment of the work behind the deliverable. Early on in the project, it was determined that the responsibility of the work behind the deliverable had been mis-assigned, and a collaborating organization had to be identified to develop the deliverable. DTU (Partner 1) was identified as the collaborating organization.

This delay has not had an impact on the work of COMPARE. In fact, the delay gave way to a more powerful result and deliverable to be shared with the EC and the public.



Background

While emerging infections by default are unpredictable, a careful review of emerging disease outbreaks has shown some pathogen classes that are more likely to (re-) emerge and spread. For instance, a recent study classified a list of 86 emerging zoonoses relevant for Europe according to 7 criteria, in order to help prioritize surveillance and preparedness planning (<http://ezips.rivm.nl/>). Algorithms for NGS/WGS/WCS -based epidemiological analysis including assessment of discriminatory power and robustness will be implemented for the main emerging pathogen classes (viral, bacterial and protozoan). These sequence-based cluster analysis tools and combined analyses of sequence data with metadata will be based on expertise with current sequence-based typing techniques (minimum spanning trees, strength and directionality of the correspondence or congruence with current reference typing methods and epidemiological cluster detection methods).

Evergreen platform

A paper describing the methodology has been submitted for publication and is under review. The article is titled 'Large scale automated phylogenomic analysis of bacterial whole-genome isolates and the Evergreen platform'.

The following is the abstract for the article:

Abstract

Public health authorities whole-genome sequence thousands of pathogenic isolates each month for microbial diagnostics and surveillance of pathogenic bacteria. The computational methods have not kept up with the deluge of data and need for real-time results.

We have therefore created a bioinformatics pipeline for rapid subtyping and continuous phylogenomic analysis of bacterial samples, suited for large-scale surveillance. To decrease the computational burden, a two level clustering strategy is employed. The data are first divided into sets by matching each isolate to a closely related reference genome. The reads then are aligned to the reference to gain a consensus sequence and SNP based genetic distance is calculated between the sequences in each set. Isolates are clustered together with a threshold of 10 SNPs. Finally, phylogenetic trees are inferred from the non-redundant sequences and the clustered isolates are placed on a clade with the cluster representative sequence. The method was benchmarked and found to be accurate in grouping outbreak strains together, while discriminating from non-outbreak strains.

The pipeline was applied in Evergreen Online, which processes publicly available sequencing data from foodborne bacterial pathogens on a daily basis, updating the phylogenetic trees as needed. It has so far placed more than 100,000 isolates into phylogenies, and has been able to keep up with the daily release of data. The trees are continuously published on <https://cge.cbs.dtu.dk/services/Evergreen>

The solution

We present a whole-genome, single nucleotide-based method for subtyping and preliminary phylogenomic analysis that circumvent the known limitations of current gene- and SNP-based approaches. PAPABAC carries out rapid and automated subtyping of bacterial whole-genome sequenced isolates and generates



COllaborative Management Platform for detection and Analyses
of (Re-) emerging and foodborne outbreaks in Europe

continuously updated phylogenetic trees based on nucleotide differences. We demonstrate two applications, a standalone version for local monitoring of bacterial isolates, and Evergreen Online, for global surveillance of foodborne bacterial pathogens. We also suggest a stable naming scheme for each isolate, making the results from the pipeline easier to communicate to others. To the best of our knowledge, no such tool exists at the moment.

The full article is available in the Annex



Annex

The submitted version of 'Large scale automated phylogenomical analysis of bacterial whole-genome isolates and the Evergreen platform' is a bioRxiv preprint first posted online Feb. 5, 2019; doi: <http://dx.doi.org/10.1101/540138>. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).

1 Large scale automated phylogenomic analysis of bacterial whole-genome isolates and the
2 Evergreen platform

3

4 Judit Szarvas¹ & Johanne Ahrenfeldt¹, Jose Luis Bellod Cisneros¹, Martin Christen Frølund Thomsen¹, Frank M.
5 Aarestrup¹, Ole Lund¹

6 ¹Research Group for Genomic Epidemiology, National Food Institute, Technical University of Denmark, Kongens
7 Lyngby, Denmark.

8

9 Abstract

10 Public health authorities whole-genome sequence thousands of pathogenic isolates each month for microbial
11 diagnostics and surveillance of pathogenic bacteria. The computational methods have not kept up with the
12 deluge of data and need for real-time results.

13 We have therefore created a bioinformatics pipeline for rapid subtyping and continuous phylogenomic analysis
14 of bacterial samples, suited for large-scale surveillance. To decrease the computational burden, a two level
15 clustering strategy is employed. The data is first divided into sets by matching each isolate to a closely related
16 reference genome. The reads then are aligned to the reference to gain a consensus sequence and SNP based
17 genetic distance is calculated between the sequences in each set. Isolates are clustered together with a
18 threshold of 10 SNPs. Finally, phylogenetic trees are inferred from the non-redundant sequences and the
19 clustered isolates are placed on a clade with the cluster representative sequence. The method was
20 benchmarked and found to be accurate in grouping outbreak strains together, while discriminating from non-
21 outbreak strains.

22 The pipeline was applied in Evergreen Online, which processes publicly available sequencing data from
23 foodborne bacterial pathogens on a daily basis, updating the phylogenetic trees as needed. It has so far placed
24 more than 100,000 isolates into phylogenies, and has been able to keep up with the daily release of data. The
25 trees are continuously published on <https://cge.cbs.dtu.dk/services/Evergreen>

26

27 Keywords

28 Phylogenomics, WGS, subtyping, SNP, automation, epidemiology, outbreak investigation

29

30

31

32 Main

33 Epidemiological typing of bacteria is used by hospitals and public health authorities, as well as animal health
34 authorities, to detect outbreaks of infectious diseases and determine trends over time. Traditionally, that
35 includes culturing and isolating the pathogen, followed by species identification and subtyping using various
36 conventional microbiological and molecular methodologies.

37 For outbreak investigation, it is necessary to place the infectious agent into a more discriminatory category
38 than species, to establish links between cases and sources. Multi-locus sequence typing (MLST) has been a
39 frequently used molecular subtyping method, where sequence types are assigned to the isolates based on the
40 combinations of alleles for 6-10 housekeeping genes¹.

41 Whole-genome sequencing (WGS) has opened a new chapter in microbial diagnostics and epidemiological
42 typing. WGS data can be used to determine both MLST types and serotype of several bacterial species^{2,3}.
43 Several studies for multiple bacterial species have shown the value of WGS for elucidating the bacterial
44 evolution and phylogeny, and identifying outbreaks⁴⁻⁶.

45 The use of WGS has enabled the unbiased comparison of samples processed in different laboratories, boosting
46 surveillance and outbreak detection, but the methods for sharing and comparing a large number of samples
47 have not been established yet^{7,8}. Therefore, a number of national, regional and international initiatives have
48 been launched with the aim of facilitating the sharing, analyses and comparison of WGS data⁹⁻¹¹.

49 Since 2012, the US Food and Drug Administration (FDA) has lead a network of public health and university
50 laboratories, called GenomeTrackr. These laboratories sequence bacterial isolates from clinical and
51 environmental samples and upload the data to the National Center for Biotechnology Information (NCBI).
52 GenomeTrackr is restricted to foodborne pathogens and currently includes data from only seven such bacterial
53 species. All raw WGS data are publicly shared, facilitating the collaboration between laboratories.¹²
54 Furthermore, the NCBI Pathogen Detection pipeline¹³ assembles the samples into draft genomes to predict the
55 nearest neighbors and construct phylogenetic trees using an exact maximum compatibility algorithm¹⁴. This
56 approach requires access to all of the raw data and very extensive computational power. In addition, no sub-
57 species taxonomical classification has been implemented at this time.

58 Focusing on the same bacterial species as GenomeTrackr, PulseNet USA also has established procedures for
59 use of WGS data for outbreak detection. In their vision, an extension of the highly successful MLST approach
60 into a core-genome (cgMLST) or whole-genome (wgMLST) scheme, with genes in the order of thousand, would
61 allow for sharing information under a common nomenclature. Meanwhile, all of the raw data could be kept
62 locally. Only data from individual strains would have to be shared when further confirmation of an outbreak is
63 required.¹¹ Consequently, a number of, at times conflicting, cg- and wgMLST schemes have been proposed for
64 a limited number of bacterial species¹⁵⁻²². Moreover, few of the proposed schemes provide a definitive
65 nomenclature of sequence types to go with the allele profiles. The existing schemes do not cover all of the
66 potential allelic variation: a recent study showed, that for *Campylobacter jejuni*, that has maintained MLST
67 schemes, only approximately 53% of the strains of animal origin could be assigned to an existing unique allelic
68 profile²³. Continuous curation of the hundreds of relevant bacterial species, that are known human, animal and
69 plant pathogens, would require great effort. A centralized database for the distribution of the allele profiles
70 and sequences would be also necessary. Furthermore, for comparable results, the same analysis pipeline or
71 software should be used for the prediction of the allelic profiles.

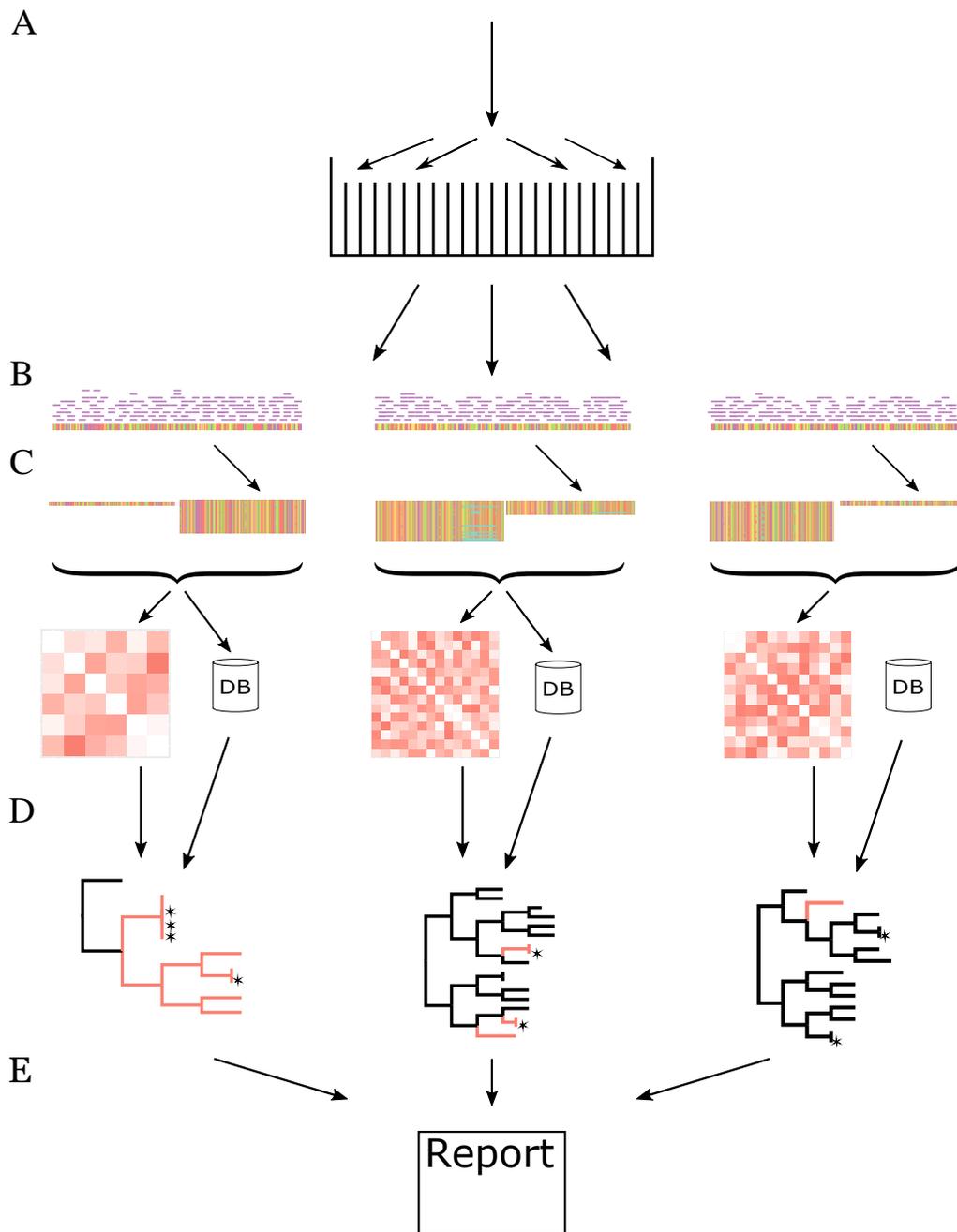
72 The approaches mentioned above yield preliminary results and, in most cases, selected WGS data are further
73 analyzed using single nucleotide profiling for outbreak detection. Here, genomic variants (single nucleotide
74 polymorphisms (SNPs), insertions and deletions) are derived by aligning WGS reads to a reference genome. For

75 each bacterial species, custom single nucleotide profiling (SNP validation, cluster threshold determination, etc.)
76 is necessary in order to achieve results that are biologically relevant and informative. The samples (of current
77 interest and historical) included in the analysis and the reference genome are chosen on a case-by-case basis,
78 usually based on subtyping results. Various SNP analysis pipelines are used by laboratories and research groups
79 for inferring phylogenetic trees for isolates of interest²⁴⁻²⁹. For example, Public Health England developed and
80 uses SnapperDB for outbreak detection without initial cluster analysis by cg- or wg-MLST. SnapperDB consists
81 of tools to create a database of SNPs compared to a given reference sequence, and assign each isolate a SNP
82 address based on single linkage clustering.³⁰

83 We present here a whole-genome, single nucleotide-based method for subtyping and preliminary
84 phylogenomic analysis, that circumvent the known limitations of current gene- and SNP-based approaches.
85 PAPABAC carries out rapid and automated subtyping of bacterial whole-genome sequenced isolates and
86 generates continuously updated phylogenetic trees based on nucleotide differences. We demonstrate two
87 applications, a standalone version for local monitoring of bacterial isolates, and Evergreen Online, for global
88 surveillance of foodborne bacterial pathogens. We also suggest a stable naming scheme for each isolate,
89 making the results from the pipeline easier to communicate to others. To the best of our knowledge, no such
90 tool exists at the moment.

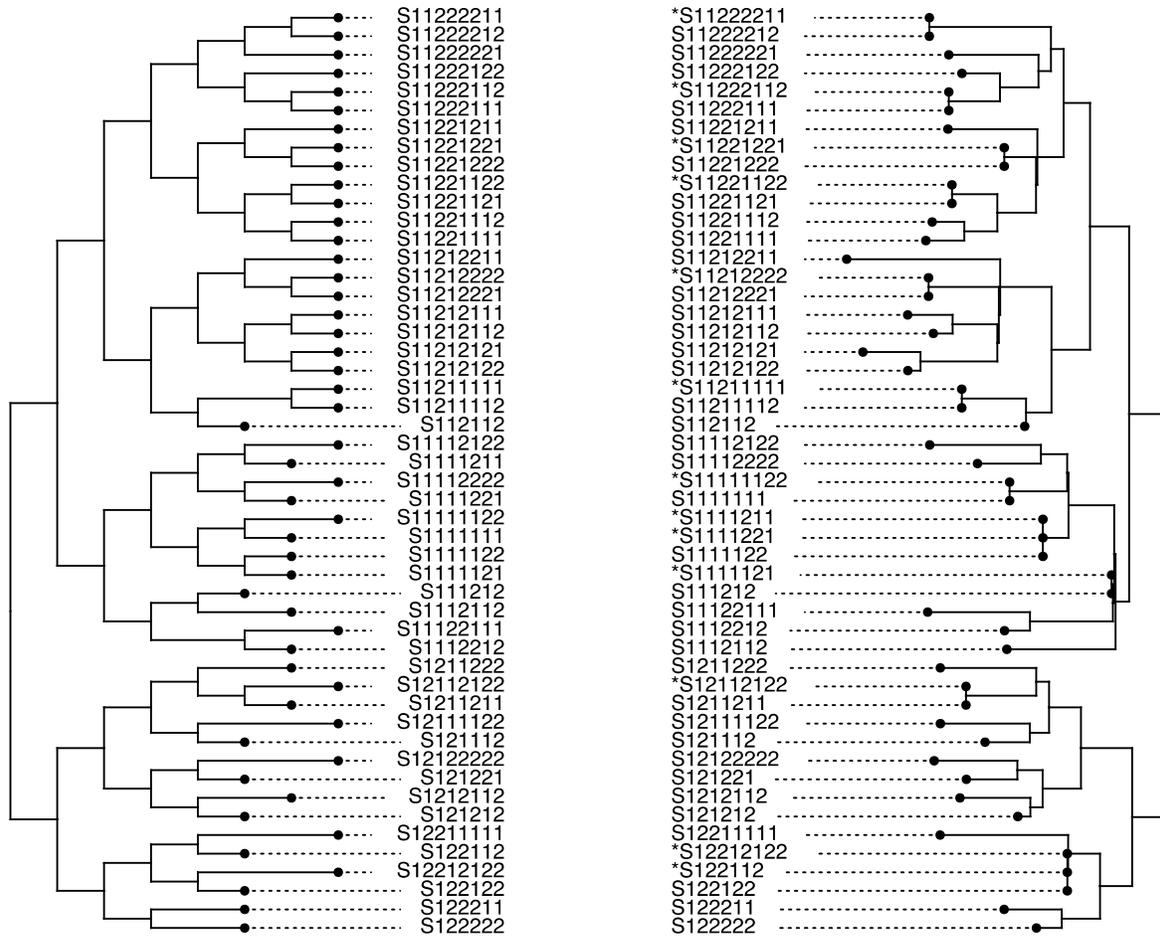
91

92



93

94 *Figure 1 Overview of PAPABAC. (A) The input raw read files are classified into sets based on k-mer similarity to NCBI RefSeq*
 95 *complete prokaryotic chromosomal genomes. (B) The raw reads are mapped to the reference genome and a consensus*
 96 *sequence is generated via strict statistical evaluation ($p < 0.05$) of the mapped bases in each position. (C) The resulting*
 97 *consensus sequences are of equal length in each template set. The new isolates in each set are clustered to the non-*
 98 *redundant isolates already in the set if the pairwise nucleotide difference based genetic distance is less than 10. The*
 99 *remaining new isolates undergo the same clustering process. (D) Pairwise genetic distance between all non-redundant*
 100 *isolate in the set is used as input for neighbor-joining algorithm. If there are less than 600 non-redundant isolates in a set,*
 101 *an approximately maximum likelihood phylogenetic tree is also inferred based on the consensus sequences (red: new*
 102 *isolates). The clustered isolates are placed back onto the trees with 0 distance to the cluster representative (marked with*
 103 *an asterisk). (E) The information about the acquired isolates, the sets, the clusters, and the phylogenetic trees is stored in*
 104 *SQLite databases, which are queried once all sets with new isolates are processed to output the results to the users.*



105

106 *Figure 2 Comparison of the ideal tree (left) to the PAPABAC maximum likelihood tree made of the in vitro experiment dataset³¹Taxa with*
107 *an asterisk were clustered together with the taxa in the same clade.*

108 Results

109 Pipeline for automated phylogenomic analysis of bacterial whole-genome sequences (PAPABAC)

110 We developed PAPABAC (Figure 1), a phylogenomic pipeline for the automated analysis of bacterial isolates,
111 that needs no additional input besides WGS data (fastq files) and generates clusters of closely related isolates.
112 PAPABAC first matches the isolates to complete bacterial chromosomal genome reference sequences with
113 greater than 99.0% sequence identity and a minimum average depth of 11. These reference sequences serve as
114 templates for the alignment of the raw reads. The aligned bases at each position are statistically evaluated to
115 determine the consensus sequence, as previously described for a nucleotide difference method³². Positions
116 that do not fulfil the significance criteria remain ambiguous, get assigned “N”, and disregarded during the
117 pairwise genetic distance calculation. These steps ensure that there is high confidence in the consensus
118 sequence that is the basis of the genetic distance estimation.

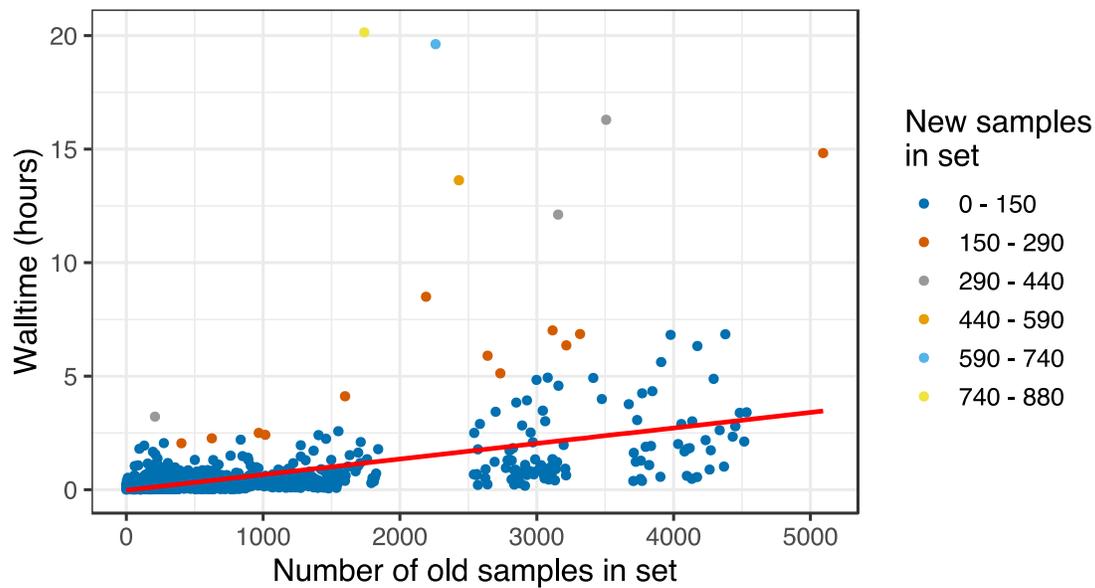
119 The pipeline retains analysis results in such a manner that input is added to the previously processed data. The
120 phylogenomic analysis is carried out on the current input and the previously found non-redundant isolates
121 (singletons and cluster representatives). The genetic distance is estimated in a pairwise manner, comparing the
122 given two sequences for all non-ambiguous positions, i.e. positions where none of the two sequences have an
123 “N” assigned. The distances between the previously processed runs are stored on disk, saving computational
124 time, and only the distances to the new isolates are computed in a given run. A clustering step during the
125 genetic distance calculation forms clusters of closely related isolates and reduces the number of similar
126 sequences in each set, and thereby also reducing the computation time. After identifying a non-redundant
127 isolate and a closely related isolate to it, the one previously deemed non-redundant will be the cluster
128 representative and kept, while the clustered one will be omitted from the subsequent runs of the pipeline.
129 However, the information about the clustering will be added to a database and the clustered isolate will be
130 placed on the inferred phylogenetic tree. The cluster representatives remain constant through the subsequent
131 runs of the pipeline, and the clusters only increase in size if new isolates are clustered with the representative.
132 Therefore, each cluster is stable and can be reliably identified by the template name and the identifier of its
133 cluster representative.

134 The pipeline can be run on a computer with 8 Gb RAM and Unix system. The computational time is reduced
135 compared to re-running the whole analysis each time new samples are added, even without parallelisation
136 (Figure S1).

137 PAPABAC was benchmarked against three SNP pipeline benchmarking datasets. An *Escherichia coli in vitro*
138 evolution experiment dataset³¹ provided 50 closely related samples on a short temporal scale with less than
139 100 nucleotide differences across the dataset. The PAPABAC maximum likelihood (Figure 2) and neighbor-
140 joining (Figure S2) trees were comparable to the ideal phylogeny of the *in vitro* experiment dataset. The
141 algorithm clustered together 7 out of 10 samples with the same ancestor that were taken on the same day and
142 presumably had less than 10 nucleotide differences between them.

143 Benchmarking against the *Campylobacter jejuni* (Figure S3A) and the *Listeria monocytogenes* (Figure S3B)
144 datasets from Timme et al.³³, PAPABAC correctly clustered the related outbreak strains (colored) and the
145 outgroups, where the genetic distance was below the clustering threshold. The topologies of the maximum
146 likelihood phylogenetic trees closely resembled the tree topologies given.

147



148

149 *Figure 3 Time requirement of the phylogenomic analysis for given number of non-redundant and new strains, on 20 CPUs.*

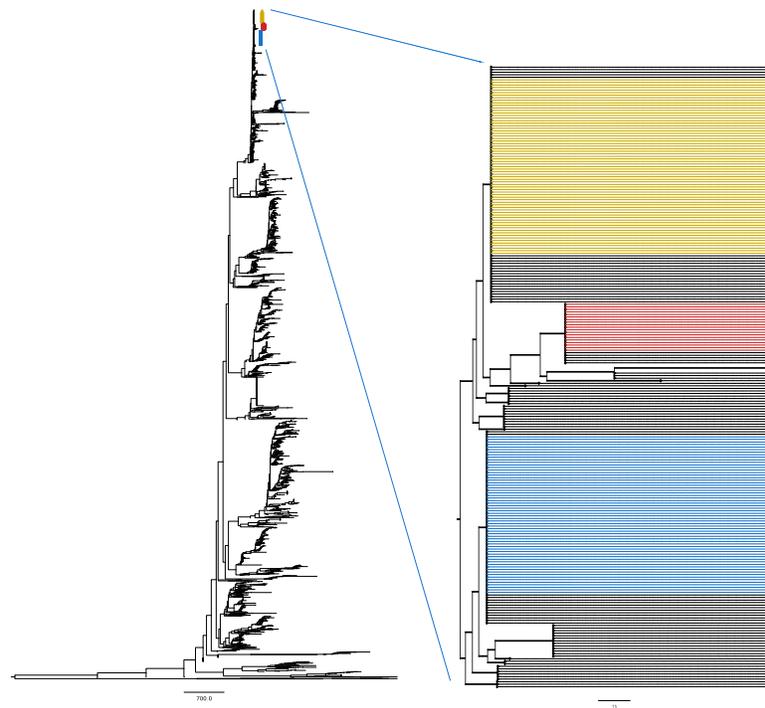
150

151 **Evergreen Online for surveillance of foodborne bacterial pathogens**

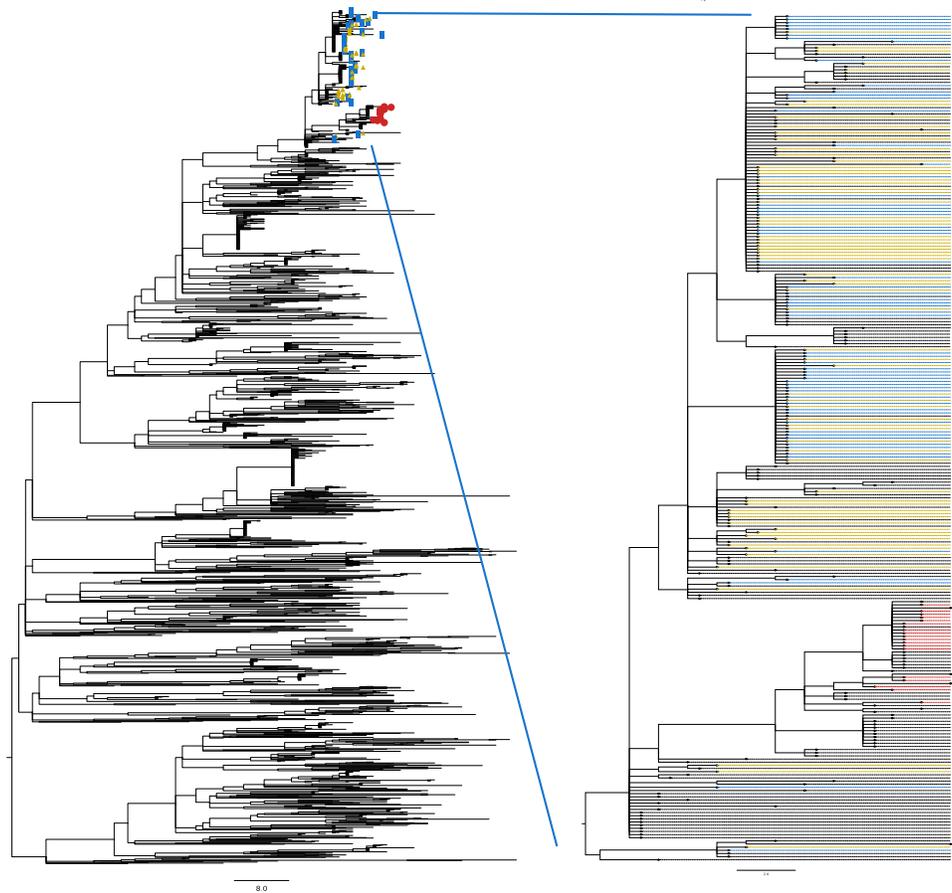
152 Evergreen Online was built on PAPABAC. Raw WGS data files of five major foodborne pathogens (*C. jejuni*, *E.*
153 *coli*, *L. monocytogenes*, *Salmonella enterica*, and *Shigella* spp.) are downloaded daily from public repositories
154 with the aim of global surveillance of potential outbreaks worldwide. The inferred phylogenetic trees and
155 information about all of the isolates in the system are available and searchable on the website
156 (<http://cge.cbs.dtu.dk/services/Evergreen>).

157 The platform has been available since October 1st 2017, with logs reliably saved since October 28th 2017. The
158 number of raw read files downloaded fluctuates with the work week of the public health laboratories. On
159 busier days, more than 800 isolates are downloaded. The average number of isolates downloaded is 418.
160 Downloading and mapping to the reference genomes take 130 minutes on average, with the majority of the
161 time spent on downloading. Alignment of the raw reads and the generation of the consensus sequences takes
162 on average 9 minutes per isolate. The computing time for the template sets is dependent on the number of
163 non-redundant and new sequences in each set, but in most cases even the slowest is finalized within five hours
164 (Figure 3).

165 As of June 26th 2018, the pipeline downloaded 82,043 isolates. Out of these, 63,276 isolates have been mapped
166 to references with at least 99.0% identity and average depth of 11 (Figure S4A). The majority of the isolates
167 were typed as *Salmonella enterica* (59.1%), followed by *Escherichia coli* (19.4%) (Figure S4B). The two largest
168 template sets are *S. Dublin* and *S. Typhimurium* serovars, with both close to 9,500 isolates in total. After the
169 homology reduction there were 3,216 and 5,093 non-redundant sequences in these sets, respectively. On
170 average, 67% of the sequences are non-redundant in the template sets, while the *E. coli* template sets are the
171 most diverse and the *Listeria monocytogenes* ones are the least diverse (Figure S4C). There were 122 isolates
172 predicted to have a type not specified by the query (Table S1). Of these, 14 isolates were mixed samples,
173 composed of both the queried and the non-queried organisms.



192



193

194 Figure 5 Selected isolates in the *Escherichia coli*_O157_H7_str_Sakai_chromosome_NC_002695_1 NJ tree (top) and on the
195 PDS000000952.271 SNP cluster maximum compatibility tree (bottom). The three largest clusters of the selected samples on the NJ tree
196 are labelled with yellow, red and blue dots. These isolates were marked with the same labels on the NCBI-PD tree. The red labelled ones
197 are on a single clade on the PD tree, while the blue and yellow isolates are mixing on two other clades.

198 Discussion

199 Whole-genome sequencing, performed alongside the traditional methods in routine microbiology, yields
200 hundreds to thousands of WGS isolates yearly in hospital, public health and food safety laboratories. This
201 amount of data is overwhelming for many, and there is a lack of methods to generate a quick overview and
202 help prioritize resources. The timely analysis of the sequencing data would allow the detection of more
203 bacterial outbreaks and aid the prevention of further spread. However, lack of human and computational
204 resources for this demanding task often hampers the prompt procession of the data. Automating the initial
205 subtyping phase would facilitate the start of an outbreak investigation. PAPABAC offers rapid subtyping for a
206 wide range of prokaryotic organisms: the supplied database covers all bacterial subtypes with complete
207 genomes present in NCBI RefSeq. Further reference genomes could be added to increase the covered sequence
208 space, but the active curation of the reference database is not required for routine usage. The selection of the
209 reference sequence for the phylogenomic analysis is fast and robust. It is independent of pre-assumptions
210 about the isolates. Misclassification during previous analysis does not introduce errors into the downstream
211 analysis. Contamination from another species is discarded during the consensus sequence generation. The
212 subtyping step via k-mer based mapping to a close reference also serves as a sequencing quality control
213 measure, because low-quality sequencing runs will typically result in isolates with low identity to any reference
214 and/or low depth. These isolates do not progress further to the phylogenomic analysis, as they would not yield
215 reliable results.

216 The phylogenomic analysis performed on the template sets has higher discriminatory power than cg- or wg-
217 MLST. The underlying nucleotide difference method was validated in five different studies^{6,31,32,34,35}. By using all
218 positions in the consensus sequences for estimating the genetic distance, instead of considering only selected
219 loci, we ensure a high level of sensitivity, as we also include mutations that occur between genes.

220 The clustering step during the genetic distance calculation was introduced in order to reduce the homology in
221 the template sets and thus reduce the computational burden as the template sets increase in size. However,
222 the clustering threshold of 10 nucleotide differences also constructs informative clusters of highly similar
223 isolates. Benchmarking with the *E. coli in vitro* evolution experiment dataset (Figure 2) showed that the
224 algorithm was capable of correctly clustering isolates that were derived from the same ancestor, while
225 distinguishing them from other closely related strains. The same sensitivity was demonstrated on empirical
226 outbreak datasets (Figure S3), where the pipeline clustered the outbreak-related strains and separated them
227 from the outgroup strains. Both the maximum likelihood inferred and the neighbor-joining trees placed the
228 outbreak strains correctly in the phylogeny. These results show, that PAPABAC provides quick and reliable
229 information about the close relatives of an outbreak strain to provide candidates to perform a more thorough
230 analysis on.

231 The design of PAPABAC means that once an isolate passed the homology reduction step, it will be present in
232 the subsequent runs of the pipeline. When an incoming isolate is highly similar to a non-redundant one, the
233 more recent will be the one that is clustered, added to the database and removed from further runs. Hence,
234 the cluster representatives and clusters are robust to the addition of new data to the analysis. Therefore,
235 PAPABAC yields a stable and communicable name for the clusters, comprised of the template name and the
236 cluster representative.

237 Evergreen Online has been steadily processing WGS data of foodborne bacterial pathogen isolates collected
238 worldwide in real time (Figure S4A). It has been able to keep pace with the flow of the generated data that
239 mainly came from public health and food safety laboratories. Excluding the download time and the optional
240 maximum likelihood based phylogenetic inference, the whole analysis is done in less than a day, even for

241 template sets with thousands of isolates (Figure 3). This turnover time facilitates quick response in a potential
 242 outbreak scenario.

243 The isolates are not distributed equally across the templates in the system (Figure S4B). Out of the five queried
 244 species, *S. enterica* isolates are disproportionately represented. Sequences in the *S. Dublin* and the *S.*
 245 *Typhimurium* LT2 template sets comprise in total approximately half of the *S. enterica* isolates. The sequence
 246 diversity in the template sets is varied, but the homology reduction on the template sets reduces the number
 247 of sequences approximately by a third, significantly decreasing the computational time. The *L. monocytogenes*
 248 template sets were the least diverse, which could be due to sampling bias: bacteria that are present in the
 249 environment are routinely sampled from food production sites multiple times, producing highly similar
 250 sequences, that are then removed from the ongoing analysis. We also tested how a large number of sequences
 251 already present in a template set would affect the ability of the pipeline to discriminate between samples
 252 (Figure 4). The template set that corresponded to the stone fruit *L. monocytogenes* outbreak dataset reference
 253 had more than 1,000 non-redundant isolates, which was ideal for the test analysis. The isolates that were part
 254 of the same outbreak clustered together and formed the two expected outbreak clusters, despite the
 255 confounding presence of the sequences already in the template set. The smaller clade, however, had a
 256 different cluster representative when using all data for the template set, compared with analysis of the
 257 outbreak data alone: an environmental sample, that could be related to the outbreak, as it was sampled from
 258 the same US state and year (California, 2014) as the samples in the outbreak dataset. These findings indicate
 259 that the pipeline is capable of identifying closely related samples, however it is necessary to conduct
 260 epidemiological analysis and apply other knowledge when interpreting the results.

261 Evergreen Online allows for automated selection of closely related isolates out of thousands, which is also the
 262 objective of NCBI-PD. *E. coli* isolates, situated on three clusters in Evergreen Online and supposedly from an
 263 outbreak, were located in NCBI-PD and their placement in the SNP cluster tree was compared to the Evergreen
 264 Online tree (Figure 5). One cluster (red) was in agreement between the two platforms, and samples from the
 265 other two (yellow and blue) clusters were intermixing on a clade on the NCBI-PD tree. The nucleotide
 266 difference counts between these samples are low and the differences between the phylogenomic methods
 267 could lead to differences in the finer details of the inferred phylogenies. The homology reducing clustering in
 268 Evergreen Online means that any sample in the cluster is less than 10 nucleotide differences from the cluster
 269 representative, however, the differences between the samples could amount to 18 nucleotides. The
 270 compatible character distances on the NCBI-PD tree between the mixed samples are less than 18 characters.
 271 Taking this into account, the observed distribution of the yellow and blue labeled samples is concordant with
 272 our results.

273 *Table 1 Comparison of pipelines for large-scale surveillance for pathogenic bacteria*

	SnapperDB	NCBI-PD	PAPABAC
For a wide range of bacterial species	X	-	X
Requires only raw sequencing reads as input	-	X	X
Whole-genome based	X	X	X
Assembly-free	X	-	X
Quality control steps	X	X	X
Automated phylogenomic analysis	-	X	X
Stable clustering of samples across runs	-	-	X
Communicable nomenclature for subtype and cluster	X	-	X
Open source	X	-	X

274

275 In summary, we developed PAPABAC with the aim of rapid subtyping and continuous phylogenomic analysis
276 on a growing number of bacterial samples. PAPABAC overcomes limitations of cg- and wg-MLST approaches by
277 tolerating genomic variation during subtyping, but providing greater sensitivity during the phylogenomic
278 analysis. It was benchmarked on datasets created for testing SNP-based pipelines, and was proved to be
279 accurate in discriminating between outbreak related and non-related samples. The software is open source and
280 fulfills expectations put to WGS-based surveillance pipelines (*Table 1*). Evergreen Online, an application made
281 for the global surveillance of foodborne bacterial pathogens, demonstrates the accuracy, speed, stability and
282 practicality of PAPABAC on thousands of samples via an on-going analysis, where the results are published
283 online.

284

285 References

- 286 1. Maiden, M. C. J. Multilocus Sequence Typing of Bacteria. *Annu. Rev. Microbiol.* **60**, 561–588 (2006).
- 287 2. Larsen, M. V. *et al.* Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria. *J. Clin. Microbiol.*
288 **50**, 1355–1361 (2012).
- 289 3. Joensen, K. G., Tetzschner, A. M. M., Iguchi, A., Aarestrup, F. M. & Scheutz, F. Rapid and easy in silico
290 serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *J. Clin. Microbiol.* **53**,
291 2410–2426 (2015).
- 292 4. Köser, C. U. *et al.* Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N.*
293 *Engl. J. Med.* **366**, 2267–75 (2012).
- 294 5. Mellmann, A. *et al.* Prospective Genomic Characterization of the German Enterohemorrhagic
295 *Escherichia coli* O104:H4 Outbreak by Rapid Next Generation Sequencing Technology. *PLoS One* **6**,
296 e22751 (2011).
- 297 6. Joensen, K. G. *et al.* Real-time whole-genome sequencing for routine typing, surveillance, and outbreak
298 detection of verotoxigenic *Escherichia coli*. *J. Clin. Microbiol.* **52**, 1501–1510 (2014).
- 299 7. WHO. *Whole genome sequencing for foodborne disease surveillance: landscape paper.* (2018).
- 300 8. Deng, X., den Bakker, H. C. & Hendriksen, R. S. Genomic Epidemiology: Whole-Genome-Sequencing-
301 Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. *Annu Rev Food Sci*
302 *Technol* **7**, 1–22 (2016).
- 303 9. GenomeTrakr Network. Available at:
304 <https://www.fda.gov/food/foodscienceresearch/wholegenomesequencingprogramwgs/ucm363134.htm>.
305 m. (Accessed: 27th June 2018)
- 306 10. COMPARE Europe. Available at: <http://www.compare-europe.eu>.
- 307 11. Nadon, C. *et al.* PulseNet International: Vision for the implementation of whole genome sequencing
308 (WGS) for global food-borne disease surveillance. *Euro Surveill.* **22**, (2017).
- 309 12. FDA, U. . Whole Genome Sequencing (WGS) Program. *U.S. Food and Drug Administration* (2016).
- 310 13. Pathogen Detection - NCBI. Available at: <https://www.ncbi.nlm.nih.gov/pathogens/>. (Accessed: 27th
311 June 2018)
- 312 14. Cherry, J. L. A practical exact maximum compatibility algorithm for reconstruction of recent evolutionary

- 313 history. *BMC Bioinformatics* **18**, 127 (2017).
- 314 15. Ghanem, M. & El-Gazzar, M. Development of *Mycoplasma synoviae* (MS) core genome multilocus
315 sequence typing (cgMLST) scheme. *Vet. Microbiol.* **218**, 84–89 (2018).
- 316 16. Higgins, P. G., Prior, K., Harmsen, D. & Seifert, H. Development and evaluation of a core genome
317 multilocus typing scheme for whole-genome sequence-based typing of *Acinetobacter baumannii*. *PLoS*
318 *One* **12**, e0179228 (2017).
- 319 17. Ghanem, M. *et al.* Core Genome Multilocus Sequence Typing: a Standardized Approach for Molecular
320 Typing of *Mycoplasma gallisepticum*. *J. Clin. Microbiol.* **56**, (2017).
- 321 18. Bletz, S., Janezic, S., Harmsen, D., Rupnik, M. & Mellmann, A. Defining and Evaluating a Core Genome
322 Multilocus Sequence Typing Scheme for Genome-Wide Typing of *Clostridium difficile*. *J. Clin. Microbiol.*
323 **56**, (2018).
- 324 19. Zhou, H., Liu, W., Qin, T., Liu, C. & Ren, H. Defining and Evaluating a Core Genome Multilocus Sequence
325 Typing Scheme for Whole-Genome Sequence-Based Typing of *Klebsiella pneumoniae*. *Front. Microbiol.*
326 **8**, (2017).
- 327 20. Cody, A. J., Bray, J. E., Jolley, K. A., McCarthy, N. D. & Maiden, M. C. J. Core Genome Multilocus
328 Sequence Typing Scheme for Stable, Comparative Analyses of *Campylobacter jejuni* and *C. coli* Human
329 Disease Isolates. *J. Clin. Microbiol.* **55**, 2086–2097 (2017).
- 330 21. Kohl, T. A. *et al.* Whole-Genome-Based *Mycobacterium tuberculosis* Surveillance: a Standardized,
331 Portable, and Expandable Approach. *J. Clin. Microbiol.* **52**, 2479–2486 (2014).
- 332 22. Moran-Gilad, J. *et al.* Design and application of a core genome multilocus sequence typing scheme for
333 investigation of Legionnaires' disease incidents. *Euro Surveill.* **20**, (2015).
- 334 23. Leekitcharoenphon, P. *et al.* Comparative genomics of quinolone-resistant and susceptible
335 *Campylobacter jejuni* of poultry origin from major poultry producing European countries (GENCAMP).
336 *EFSA Support. Publ.* **15**, (2018).
- 337 24. Kvistholm Jensen, A. *et al.* Whole-genome Sequencing Used to Investigate a Nationwide Outbreak of
338 Listeriosis Caused by Ready-to-eat Delicatessen Meat, Denmark, 2014. *Clin. Infect. Dis.* **63**, 64–70 (2016).
- 339 25. Schjørring, S. *et al.* Cross-border outbreak of listeriosis caused by cold-smoked salmon, revealed by
340 integrated surveillance and whole genome sequencing (WGS), Denmark and France, 2015 to 2017. *Euro*
341 *Surveill.* **22**, (2017).
- 342 26. Ford, L. *et al.* Incorporating Whole-Genome Sequencing into Public Health Surveillance: Lessons from
343 Prospective Sequencing of *Salmonella* Typhimurium in Australia. *Foodborne Pathog. Dis.* **15**, 161–167
344 (2018).
- 345 27. Holmes, A., Dallman, T. J., Shabaan, S., Hanson, M. & Allison, L. Validation of Whole-Genome
346 Sequencing for Identification and Characterization of Shiga Toxin-Producing *Escherichia coli* To Produce
347 Standardized Data To Enable Data Sharing. *J. Clin. Microbiol.* **56**, (2018).
- 348 28. Woksepp, H., Ryberg, A., Berglind, L., Schön, T. & Söderman, J. Epidemiological characterization of a
349 nosocomial outbreak of extended spectrum β -lactamase *Escherichia coli* ST-131 confirms the clinical
350 value of core genome multilocus sequence typing. *APMIS* **125**, 1117–1124 (2017).
- 351 29. Davis, S. *et al.* CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-
352 generation sequence data. *PeerJ Comput. Sci.* **1**, e20 (2015).

- 353 30. Dallman, T. *et al.* SnapperDB: a database solution for routine sequencing analysis of bacterial isolates.
354 *Bioinformatics* **81**, 3946–3952 (2018).
- 355 31. Ahrenfeldt, J. *et al.* Bacterial whole genome-based phylogeny: construction of a new benchmarking
356 dataset and assessment of some existing methods. *BMC Genomics* **18**, 19 (2017).
- 357 32. Leekitcharoenphon, P., Nielsen, E. M., Kaas, R. S., Lund, O. & Aarestrup, F. M. Evaluation of Whole
358 Genome Sequencing for Outbreak Detection of *Salmonella enterica*. *PLoS One* **9**, e87991 (2014).
- 359 33. Timme, R. E. *et al.* Benchmark datasets for phylogenomic pipeline validation, applications for foodborne
360 pathogen surveillance. *PeerJ* **5**, e3893 (2017).
- 361 34. Kaas, R. S., Leekitcharoenphon, P., Aarestrup, F. M. & Lund, O. Solving the problem of comparing whole
362 bacterial genomes across different sequencing platforms. *PLoS One* **9**, e104984 (2014).
- 363 35. Joensen, K. G. *et al.* Evaluating next-generation sequencing for direct clinical diagnostics in diarrhoeal
364 disease. *Eur. J. Clin. Microbiol. Infect. Dis.* **36**, 1325–1338 (2017).
- 365 36. Clausen, P. T. L. C., Aarestrup, F. M. & Lund, O. Rapid and precise alignment of raw reads against
366 redundant databases with KMA. *BMC Bioinformatics* **19**, 307 (2018).
- 367 37. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. Selection of representative protein data sets. *Protein*
368 *Sci.* **1**, 409–417 (1992).
- 369 38. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees.
370 *Mol. Biol. Evol.* **4**, 406–25 (1987).
- 371 39. Studier, J. & Keppler, K. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol.* **5**,
372 729–731 (1988).
- 373 40. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic
374 algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–74 (2015).
- 375 41. Huerta-Cepas, J. *et al.* ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol.*
376 *Biol. Evol.* **33**, 1635–1638 (2016).
- 377 42. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods*
378 *Ecol. Evol.* **3**, 217–223 (2012).
- 379 43. CDC. Multistate Outbreak of *E. coli* O157:H7 Infections Linked to Romaine Lettuce (Final Update) |
380 Investigation Notice: Multistate Outbreak of *E. coli* O157:H7 Infections April 2018 | *E. coli* | CDC.
381 Available at: <https://www.cdc.gov/ecoli/2018/o157h7-04-18/index.html>. (Accessed: 7th August 2018)
- 382
- 383

384 **Methods**

385 **Bioinformatics pipeline: PAPABAC**

386 The pipeline takes raw whole-sequencing reads (fastq files) as input. Matching reference sequences
387 (templates) in our reference database, that have greater than 99.0% identity, are identified for the isolates
388 using 16-mers via KMA³⁶ in sparse mode. Multiple templates are accepted, if they meet the criteria, allowing
389 for the procession of mixed samples. Information about the runs and their templates are inserted into the main
390 SQLite database. The isolates are grouped into sets according to the matched templates. The next steps are
391 performed in these sets in parallel. The isolate reads are mapped to the template using the mapping algorithm

392 of NDtree³², yielding equal-length consensus sequences. The Z-score threshold for accepting a base is set to
393 1.96.

394 Genetic distance based on nucleotide difference is calculated pairwise between the previous, non-redundant
395 isolates and the new isolates. Positions with ambiguous bases are discarded. The new isolates are clustered to
396 the non-redundant ones with a threshold of 10, in order to reduce the homology in each set and form
397 informative clusters. In this step, the non-redundant isolate is prioritized over the new isolate and becomes the
398 cluster representative. After the clustering, the remaining new isolates are clustered together with the
399 Hobohm 1 algorithm³⁷. In this case, the cluster representative is the one that has already passed the
400 redundancy threshold. The information about new or extended clusters is saved to the main SQLite database. A
401 distance matrix is constructed for all non-redundant isolates and saved to disk for use in the next run. A
402 distance-based phylogenetic tree is inferred by neighbor-joining^{38,39}. If there are less than 600 non-redundant
403 isolates in the set, then a whole-genome based approximate maximum likelihood phylogenetic tree is also
404 inferred using IQ-tree⁴⁰, where the neighbor-joining tree is the starting tree and the GTR nucleotide
405 substitution model is used. The clustered isolates are placed back onto the clades with zero distances to the
406 cluster representative. Their tip labels start with an asterisk. The information about the trees is saved to the
407 main SQLite database.

408 When all the phylogenetic trees with new isolates have been inferred, then the main SQLite database is
409 queried for the list of all isolates, their templates, cluster representatives (if there is any) and the latest
410 phylogenetic tree they are on. This information is printed to a tab-separated file.

411 Scripts and installation instructions are available on bitbucket:

412 <https://bitbucket.org/genomicepidemiology/evergreen>

413 **Online Evergreen platform**

414 A query is made to the National Center for Biotechnology Information (NCBI) Sequencing Read Archive (SRA)
415 for the newly published Illumina paired-end sequenced isolates of *Campylobacter jejuni*, *Escherichia coli*,
416 *Listeria monocytogenes*, *Salmonella enterica*, and *Shigella spp.* on a daily basis. Fastq files of raw sequencing
417 reads and the corresponding metadata (collection date, location, institute, source, etc.) are acquired either
418 from SRA or from the European Nucleotide Archive (ENA). The downloaded isolates are the input to PAPABAC.
419 The metadata are saved in the main SQLite database, and added to the tip labels on the phylogenetic trees.

420 Once all instances of the second wrapper script have finished, then the SQLite databases are queried for the list
421 of available phylogenetic trees (the maximum likelihood trees preferred over neighbor-joining ones), changes
422 in the clusters and the list of all isolates in the system, which is then used to update the website.

423 **Architecture**

424 The pipeline is written in Python 2.7 and Bash in Unix environment. In addition to the standard Anaconda
425 Python 2.7 packages, it also requires ETE Toolkit v3.0⁴¹ and Joblib v0.11 (<https://pythonhosted.org/joblib>)
426 packages to be installed. Neighbor program from the PHYLIP package v3.697
427 (<http://evolution.genetics.washington.edu/phylip.html>) and IQ-tree v6.0⁴⁰ are used for the phylogenetic tree
428 inference. The SQL database management is performed with SQLite v3.20.1 (<https://www.sqlite.org>).

429 The two main parts of the pipeline have their own wrapper scripts. PAPABAC can be run on a personal
430 computer with as few as four cores.

431 Evergreen Online is running on a high-performance computing cluster, utilizing the Torque (Adaptive
432 Computing Inc., USA) job scheduler. The first wrapper is run in one instance on 20 cores, meanwhile the second

433 wrapper is run once on 20 cores for each template that has at least one new run, in a parallel fashion. When all
434 of these instances are finished running, a Bash script is launched to collect the information from the SQL
435 database, the website is updated and the job for the next day is scheduled.

436 **Reference database**

437 The reference sequences are complete prokaryotic chromosomal genomes from the NCBI RefSeq database.
438 Homology reduction was performed at a 99.0% sequence identity threshold with the Hobohm 1 algorithm. The
439 curated NCBI prokaryotic reference genomes were given priority in the process. The reference sequences and
440 the classification database could be downloaded via ftp
441 (<ftp://ftp.cbs.dtu.dk/public/CGE/databases/Evergreen/>).

442 **Website**

443 The phylogenetic trees are interactively visualized on the website (<https://cge.cbs.dtu.dk/services/Evergreen/>)
444 using the PhyloCanvas API (<http://phylocanvas.org>). The isolates and clusters can be searched by SRA run ID,
445 which allows the quick localization of the clusters that increased in size via their cluster representative.

446 **Computational time comparison of continued phylogenomic analysis**

447 101 samples from the *Escherichia coli* in vitro evolution experiment dataset by Ahrenfeldt et al. were batched
448 according to their sampling time. The parallelization in PAPABAC was disabled. The traditional method meant
449 that the analysis was carried out on all the samples up to the given batch, starting anew each time, but using
450 the same scripts as PAPABAC.

451 **Benchmarking of PAPABAC with the *Escherichia coli* in vitro evolution experiment dataset by Ahrenfeldt et al.**

453 The last samples in each lineage were selected for the benchmarking. Therefore, the benchmarking dataset
454 constituted 50 tips on the ideal phylogeny. These samples were batched according to their sampling time (6th,
455 7th and 8th day). The batches were processed by PAPABAC chronologically. The pipeline was run with the
456 default parameters. Both maximum likelihood and neighbor-joining trees were inferred.

457 The phylogenetic trees inferred on all 50 isolates were trimmed for the reference sequence and compared with
458 the ideal phylogeny using the phytools R package⁴².

459 **Benchmarking of PAPABAC with datasets from Timme et al.**

460 Each dataset was downloaded with the provided script into a distinct directory. The pipeline was run
461 individually on the datasets with default parameters. If the isolates were mapped to more than one template,
462 the phylogenetic trees of the template set with the highest number of isolates were evaluated. The maximum
463 likelihood trees were visually compared to the ideal phylogenies and checked for the distribution of the isolates
464 amongst the clades.

465 **Comparison with the NCBI Pathogen Detection platform**

466 *Escherichia coli* isolates were queried from the SQL database of Evergreen Online (EO) for the period of 2018-
467 03-15 and 2018-06-01, corresponding to a multistate outbreak of *E. coli* O157:H7 in the USA⁴³. These samples
468 were subtyped using traditional MLST², as it was assumed, that the sequence type with the most isolates would
469 also include the outbreak samples. Sequence type 11, which is commonly corresponds to the O157:H7
470 serotype, was selected for further analysis. The corresponding samples and their SNP clusters were found in
471 the NCBI-PD platform. The phylogenetic tree for the SNP cluster with the most samples (PDS000000952.271)

472 was downloaded. The common samples were marked on both the NCBI-PD and the EO phylogenetic tree
473 (*Escherichia_coli_O157_H7_str_Sakai_chromosome_NC_002695_1*). The marked samples on the three biggest
474 clusters on the EO tree were labeled, and their placement on the NCBI-PD tree was visually inspected.

475 Supplementary material

476

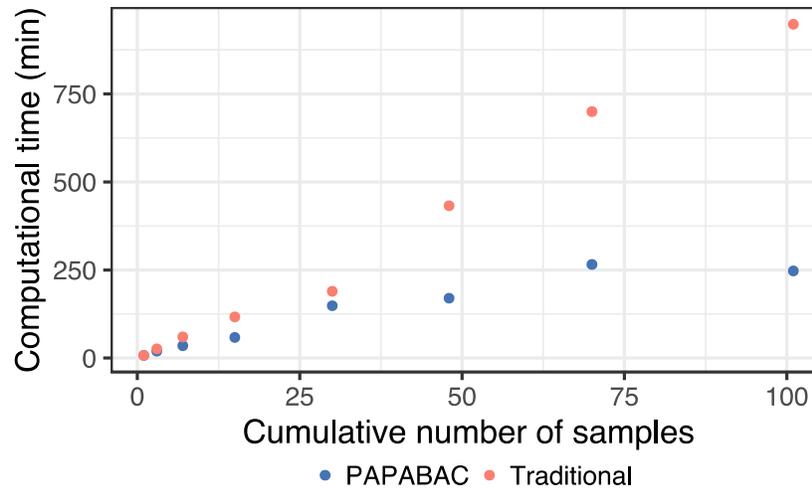
477

Table S1 Non-queried species, due to mislabelled or mixed samples

Genus	Species	Isolate
<i>Bacillus</i>	<i>subtilis</i>	3
<i>Bacillus</i>	<i>pumilus</i>	2
<i>Campylobacter</i>	<i>coli</i>	58
<i>Campylobacter</i>	<i>fetus</i>	1
<i>Citrobacter</i>	<i>amalonaticus</i>	1
<i>Enterobacter</i>	<i>cloacae</i>	2
<i>Enterococcus</i>	<i>faecalis</i>	1
<i>Escherichia</i>	<i>albertii</i>	5
<i>Hafnia</i>	<i>alvei</i>	3
<i>Klebsiella</i>	<i>pneumoniae</i>	7
<i>Listeria</i>	<i>ivanovii</i>	1
<i>Morganella</i>	<i>morganii</i>	7
<i>Peptoclostridium</i>	<i>difficile</i>	1
<i>Proteus</i>	<i>mirabilis</i>	7
<i>Providencia</i>	<i>stuartii</i>	2
<i>Pseudomonas</i>	<i>aeruginosa</i>	6
<i>Raoultella</i>	<i>ornithinolytica</i>	1
<i>Salmonella</i>	<i>bongori</i>	11
<i>Staphylococcus</i>	<i>epidermidis</i>	1
<i>Streptococcus</i>	<i>agalactiae</i>	1

478

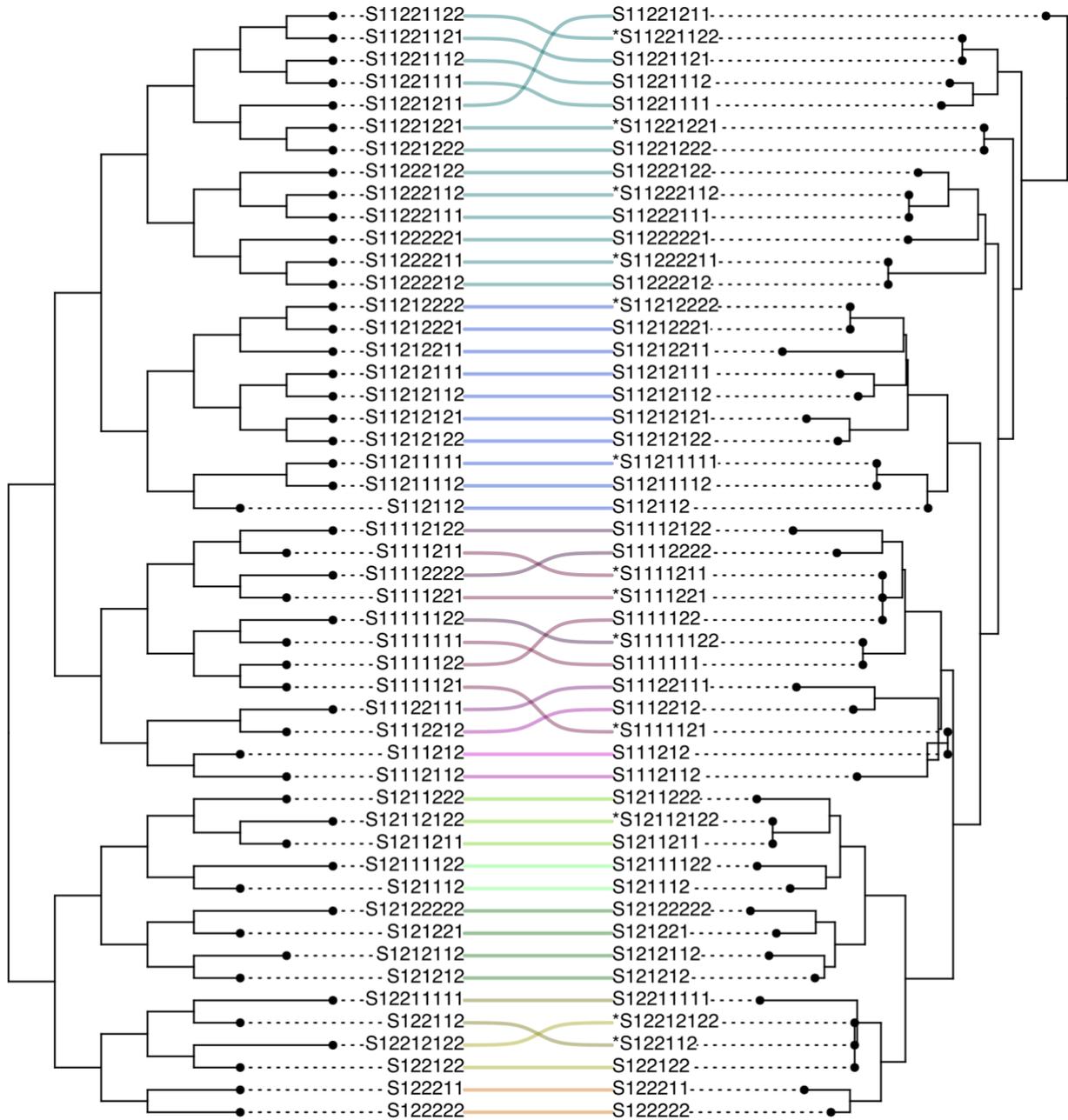
479



480

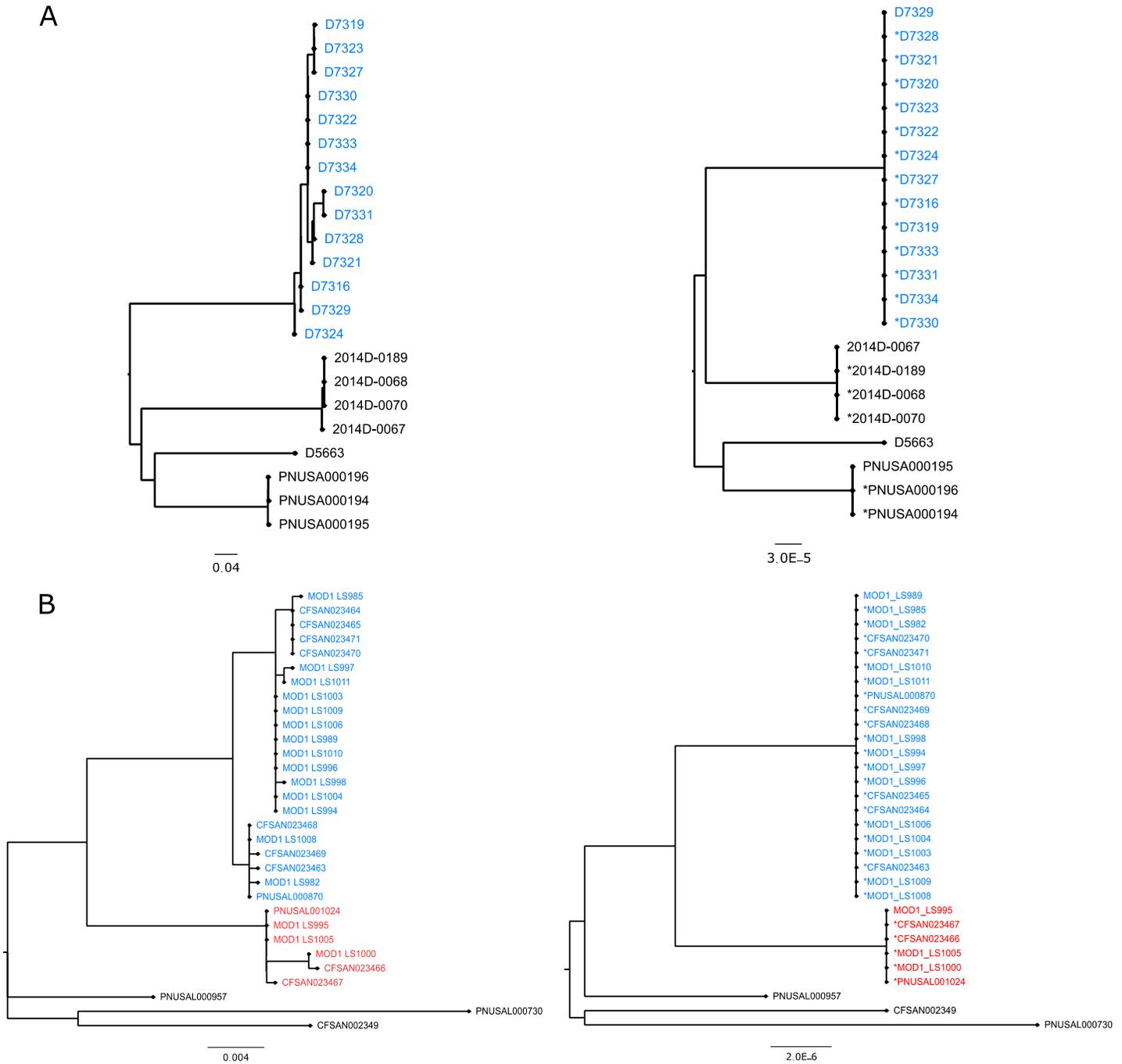
481 Figure S1 Computational time of the *Escherichia coli* in vitro evolution dataset where the samples were added in batches based on the
482 sampling time.

483



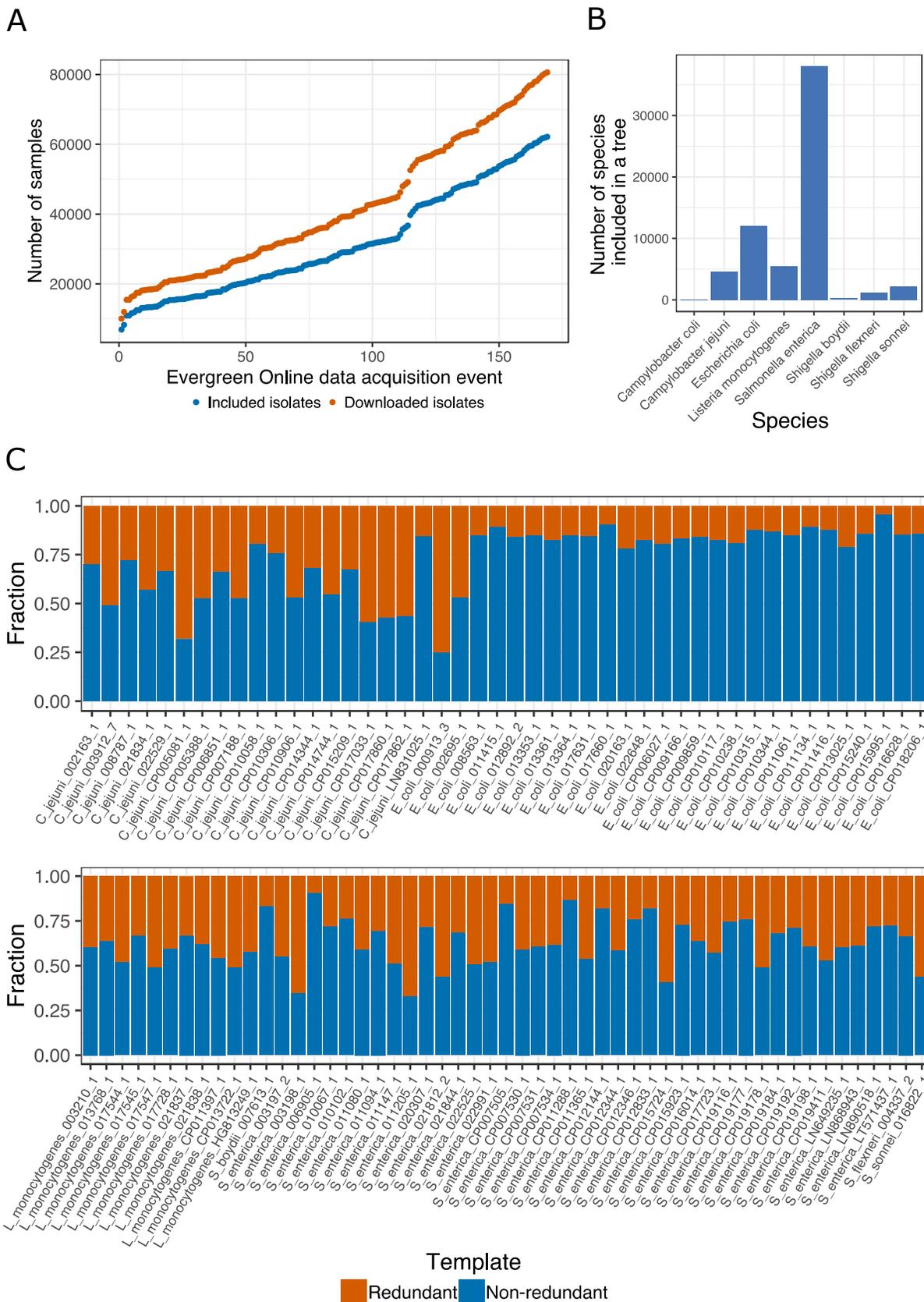
484

485 *Figure S2 Comparison of the ideal tree (left) to the PAPABAC neighbor-joining tree made of the in vitro experiment dataset³¹Taxa with an*
486 *asterisk were clustered together with the taxa in the same clade.*



487

488 *Figure S3 Maximum likelihood trees of (A) Campylobacter jejuni and (B) Listeria monocytogenes SNP pipeline benchmarking datasets.*
 489 *The trees on the left are the "ideal" phylogenies by Timme et al. The colored (blue, red) clades contain the outbreak strains, while the*
 490 *black ones are non-related isolates. The reference sequences were trimmed from the trees.*



491

492 Figure S4 A) Number of downloaded and included isolates as function of data acquisition events B) Number of isolates for the species we
 493 query for C) Fraction of non-redundant isolates in template sets larger than 100 isolates