SUPPLEMENTARY MATERIAL

Refinement of whole-genome multilocus sequence typing analysis by addressing gene paralogy

Ji Zhang[1#], Jani Halkilahti[2], Marja-Liisa Hänninen[1] and Mirko Rossi[1#]

[1]Department of Food Hygiene and Environmental Health, University of Helsinki, Helsinki, Finland

[2]National Institute for Health and Welfare, Helsinki, Finland

[#]Corresponding authors: Mirko Rossi and Ji Zhang

mirko.rossi@helsinki.fi

ji.zhang@helsinki.fi

OVERVIEW OF THE GENOME PROFILER (GEP)

GeP was written in PERL (http://www.perl.org) and tested with Perl v5.16.2 in Mac OS X 10.9.5 and Ubuntu Linux 12.04 LTS operating systems. The program is freely available at (http://sourceforge.net/projects/genomeprofiler/) under the terms of the GNU General Public License. GeP depends on BLAST+ (1) to perform sequence search and MAFFT (2) to align allele sequences. A flow chart of GeP logic is shown in Figure S1. GeP requires three input files: 1) the complete or draft genomic sequences of the isolates in FASTA format; 2) a text file listing the names of the sequence files of the isolates; and 3) an annotated reference genome sequence in GenBank format. In case of running with the option "-o", the entire 'scheme' folder that created by GeP in the *ad hoc* mode (see blow) is required.

AD HOC WHOLE-GENOME MLST RUNNING MODE

GeP initiates the analysis by parsing the GenBank file of the reference genome. For each locus, the amino acid sequences and their coding sequences are extracted and written separately to single-FASTA files. The application MAKEBLASTDB in the BLAST+ package is called to format these sequence files to BLAST databases. From the reference genome, gene information, such as locus tag, product, gene length and length of all intergenic regions, are recorded and used for the analysis. The allele number of each locus of the reference genome is set to “1”. The analysis continues by BLAST searching each genome sequence against the reference sequence databases after turning off the “-dust” (BLASTN) or “-seg” option (BLASTX) to prevent filtration of low complexity and repetitive regions. If the query genome is in the form of multiple contig sequences, before performing the BLAST, GeP concatenates the contigs into a single sequence using a spacer formed by a fixed number of Ns (default 20,000 bp).

GeP starts the analysis by calling BLASTN to search the nucleotide sequence reference database using first the reference, and then one by one the other test genome sequences as query. By default, all hits that align to less than 50% length of the locus (coverage < 50%) and have nucleotide

identity less than 80% are discarded. All the remaining hits are 'valid hits'. The program will first search for possible multi-copy gene in the reference genome (self-blasting). If there is more than one valid hit of the gene were found in the reference, the gene would be marked for the following analysis. For the test genomes, if there is only one valid hit, it would be deemed as the correct ortholog in the query genome unless the gene was marked as multi-copy gene in the self-blasting step (for multi-copy gene, see below). Based on the nucleotide identity within the reference locus database, an allele number will be attributed to the located ortholog only if the hit covers the entire length of the reference locus (coverage=100%). Otherwise the locus will be marked as 'T' (Truncated) and excluded from the analysis. If the located ortholog is not identical to any of the sequences in the reference locus database (nucleotide identity < 100%), a new allele number will be assigned. Then, the new allele sequence will be extracted from the genome and added to the reference locus nucleotide database. After assigning the allele number, GeP will move to the following locus.

If the above procedure failed to locate the ortholog in the query genome, GeP will call BLASTX to translate the genome sequence and search corresponding amino acid sequence databases for the reference locus. If a single BLAST hit covers 50% or more of the full length of one sequence of the reference locus database and has 80% or more amino acid identity, it will be identified as the correct ortholog in the query genomic sequence (for multi-copy gene, see below). The nucleotide sequence of the located ortholog will be extracted from the genome only if it covers the entire length of the reference locus (coverage=100%). A new allele number will then be assigned and added to the reference allele nucleotide database. Otherwise, the locus will be marked as 'T' and excluded from the analysis. If BLASTX also failed to locate the ortholog gene in the query genome, the locus will be marked as 'M' (Missing) and excluded from the analysis.

Multi-copy genes

GeP separates orthologs from paralogs by looking for CGN. It assumes that the contiguity and the distance of any given two neighboring genes should be conserved between the reference genome and the tested genomes. If multiple copies of a gene were found in the genomes, the pair that follows the CGN is likely to be orthologs. GeP records the length of all intergenic regions in the reference genome. Then, to allow wobbling, it defines a value for the 'expected distance to the previous locus' (expected d) by adding 10 extra base pairs to the intergenic region value or, in the case of gene overlapping, by giving a fixed value of -1. If multiple valid BLASTN or BLASTX hits (coverage>=50% and identity>=80% by default) for a given locus are found in the query or reference genomes, GeP treats the hits as potential orthologs only when they are located inside the range of 'expected d', and the program will automatically select the one with the smallest d value. If none of the valid BLAST hits were within the range of 'expected d', the locus in this genome will be marked as 'D' (Duplicated) and excluded from the analysis.

Extraction of new allele sequences

The new allele sequence is extracted from the study genome sequences based on the coordinates of the BLAST alignments. If the coordinates are defined by BLASTX, GeP will add three nucleotides of the downstream to the end of the extracted allele sequence (e.g., to include the stop codon). GeP also checks if the extracted sequences contain nucleotide ambiguity. If ambiguity is found, the locus will be marked as 'N' (Nucleotide ambiguity), and it will not be counted as a new allele and will be excluded from the analysis.

Summary of the results

After locating all of the loci in the query genomes and assigning the corresponding allele number, GeP will summarize the genetic differences of all shared-loci (the loci having allele information in all of the queries) and write the results to the following files: (i) *output.txt*, (ii) *difference_matrix.html*, (iii) *Splitstree.nex*, (iv) *allele_profile.txt*, and (v) two core genome files *clonalframe.dat* and *core_genomes.fas*. The *output.txt* file records the information of all the loci in

each of the test genome sequences. All the other output files are derived from *output.txt*. The HTML output file contains a summary of the analysis and a matrix of pairwise differences between the allelic profiles of the samples. The numbers in the matrix are hyperlinks, which allow the user to view a detailed list of the genes with different allelic assignments in the pairwise comparison. By clicking the name of the gene, the user can visually inspect the sequence alignments and identify all of the genetic differences in the selected locus. The file *Splitstree.nex* includes the allele profile of the isolates in NEXUS format, which can be opened in Splitstree 4 (3) and visualized either using split decomposition or neighbor-net algorithm (4). The *allele_profile.txt* is a tab-delimited format of the allele profiles of the isolates, which can be used in downstream population structure analysis programs, such as STRUCTURE (5) or BAPS (6). Finally the two aligned core genome files *clonalframe.dat* (eXtended Multi-Fasta format) and *core_genomes.fas* (Fasta format) can be used as input for whole-genome evolution and recombination analysis programs.

FIXED SCHEME MLST RUNNING MODE

After the first run in *ad hoc* mode, the wgMLST scheme built in the analysis will be saved to files in the 'scheme' folder in the working directory. New isolates can be analyzed against the wgMLST scheme by using the option "-o", and the new allele information will be added to the existing scheme. This option allows the users to expand the allele database of a peculiar set of isolates and to develop a specific nomenclature, which can be used for follow-up epidemiological studies. Also it makes easy to share and transfer the wgMLST scheme between labs, upon which a standardized wgMLST scheme can be built.

BENCHMARK DATA SETS

To test our program and demonstrate its capabilities, a collection of 19 *Campylobacter jejuni* ST-45 isolates was used in this study. Ten of the isolates originated from three independent waterborne outbreaks that occurred in 2000 and 2001 in Finland (7). One of the isolates was obtained from a tap water sample, and the others were isolated from patients (Table S1). According to Finnish

125 legislation, no ethical approval is needed for public health response to a waterborne outbreak. The
126 other nine non-outbreak-associated isolates were obtained from four Finnish chicken farms. They
127 were either isolated at a slaughterhouse in summer 2012 (farm A, B, and C) or fecal and
128 environmental samples of a farm in 2003 (farm D). The collection thus included both
129 epidemiologically associated and non-associated isolates (Table 1), and they all had similar KpnI
130 patterns (7,8) (and unpublished data).
131 In addition to the seven previously sequenced isolates (4031, IHV116260, IHV116292, 6237, 6236,
132 6538 and 6497) (EMBL project number PRJEB4165) (8), the other 12 isolates were sequenced by
133 Illumina sequencing technology with 100 cycles paired-end reads and Nextera XT library
134 preparation. Sequencing was performed at the Institute for Molecular Medicine Finland. All the
135 genomic sequences were assembled using SPAdes genome assembler version 3.1.1 (9).
136 The wgMLST analyses of the 19 *C. jejuni* WGS data were performed with GeP, BIGSdb Genome
137 Comparator (10) and SeqSphere+ version 1.0 (Ridom GmbH, Münster, Germany) (11) using the
138 annotated genome sequence of *C. jejuni* 4031 (GenBank Acc. NC_022529) (8) as reference. The
139 default settings were used in all three programs. The online tool 'Genome Comparator' hosted by
140 the PubMLST website (http://pubmlst.org/campylobacter/, accessed 17.10.2014) was used to test
141 the performance of the BIGSdb Genome Comparator. All assembled data were deposited in the
142 PubMLST database (10), and the accession numbers are listed in Table S1.
143 **RESULTS AND DISCUSSION**
144 **Split-decomposition**
145 The topologies of the splitgraph generated by GeP and SeqSphere+ seemed identical (Fig. S2a; Fig.
146 S2b) and similar to the one produced by BIGSdb GC (Fig. S2c). The results from both GeP and
147 SeqSphere+ revealed that, except for outbreak 1, the core genomes of *C. jejuni* belonging to same
148 outbreak or isolated within the same farm were highly similar. In contrast, the isolates between the
149 outbreaks or farms were separated from each other, indicating that all analysis tools were able to

150 separate epidemiologically associated isolates from non-associated isolates, confirming the results
151 of our previous studies (8, 12). Overall, the results of BIGSdb GC overlapped the results produced
152 by the other two programs, with the exception of a visible net-like structure separating the isolates
153 of farm B (Fig. S2c). After removing the missing allele state in the BIGSdb GC allele profile (see
154 main text) the topology of the splitgraph resembles exactly the ones produced by GeP and
155 SeqSphere+.
156 Despite the general similarity in the splitgraphs, the numbers of identical and polymorphic
157 shared-loci found by the three programs were different, which affected pairwise allelic differences
158 of the isolates. In fact, the average of the intra-cluster allele differences calculated by GeP, BIGS
159 GC and SeqSphere+ was 3.3, 11.3 and 1.2, respectively.
160 **Failing to choose the orthologous gene from the paralogous gene (Error type I)**
161 GeP found in 306 cases, 34 loci containing possible paralogous genes in the tested genomes (Table
162 S2). GeP was able to use CGN to differentiate orthologs from the paralogs in 222 of these cases
163 (DATASET S3). Among the 34 loci, four (BN867_00630, BN867_00640, BN867_06950 and
164 BN867_09650) were able to be used to generate allele profiles, six (BN867_05110, BN867_05120,
165 BN867_06960, BN867_09580, BN867_09590 and BN867_09640) were excluded solely because of
166 inconsistent gene synteny and the other 24 loci were excluded because of either missing, truncation
167 or nucleotide ambiguity. SeqSphere+ excluded 31 out of 34 of these loci from the analysis, failing
168 in the identification of the duplication in several cases, which resulted in the omission of one locus
169 (Table 4). For ten of these loci (Table 2 and 4), SeqSphere+ did not report any information,
170 presumably because they are duplicated in the reference genome and they were excluded from the
171 original list of reference loci. BIGSdb GC was more prone to error type I by including in the
172 analysis 15 loci excluded by GeP, 4 of which were omitted by the latter solely due to duplication
173 (BN867_05110, BN867_05120, BN867_09580 and BN867_09590). BIGSdb GC mistakenly
174 identified these four loci as identical (Table 2 and 4). In addition, BIGSdb GC tagged

175 BN867_14900 as a paralogous locus, but no extra copy of this locus was found among the 19 tested
176 genome sequences either by GeP or SeqSphere+.
177 **Homopolymeric tracts (Error type VI)**
178 SeqSphere+ wrongly excluded from the analysis all loci containing homopolymeric tracts, which is
179 commonly observed in *C. jejuni* genomes (12, 13), if the length of the tracts differs from the
180 reference genome. GeP takes homopolymeric tracts of different length into account by assigning
181 different allele numbers. The user can later easily inspect the sequence alignment in the GeP output
182 files and make a decision whether to include these loci. Variations at hypervariable regions might
183 cause overestimation in the allele differences, but this phenomenon is usually limited to a small
184 portion of the core genome (approximately 1-2% of shared-loci in *C. jejuni*) and appears to be
185 relevant only when highly similar genomes are compared (e.g., different isolates of the same
186 outbreak) (12, 14). For example, we recently showed that variation in the lengths of homopolymeric
187 tracts are the only differences detected in outbreak-associated *C. jejuni* isolates, but they have been
188 considered irrelevant in defining relatedness between the isolates (12).
189

190 **TABLES**

191 **Table S1** List of 19 *C. jejuni* ST-45 isolates used in this study

Origin	Isolate	Source	Year	Accession n	Reference
Waterborne Outbreak 1	4031	Water	2000	2692	8, 7
	IHV116260	Human	2000	2693	8, 7
	IHV116292	Human	2000	2694	8, 7
Waterborne Outbreak 2	540	Human	2001	2695	7
	543	Human	2001	2697	7
	544	Human	2001	2698	7
Waterborne Outbreak 3	T-71726	Human	2001	2699	7
	T-71727	Human	2001	2700	7
	T-71731	Human	2001	2701	7
	T-71732	Human	2001	2702	7
Chicken farm A	6237	Chicken feces	2012	2684	8
	6236	Chicken feces	2012	2685	8
Chicken farm B	6538	Chicken feces	2012	2686	8
	6541	Chicken feces	2012	2689	This study
Chicken farm C	6498	Chicken feces	2012	2690	This study
	6497	Chicken feces	2012	2691	8
Chicken farm D	4028	Farm environment	2003	2681	This study
	4947	Farm environment	2003	2682	This study
	4948	Farm environment	2003	2683	This study

192

193

Table S2. Allele state of 34 putative multi-copy loci found by GeP in each isolate assigned by GeP, BIGSdb GC and SeqSphere+. The isolates are in the following order (from right to left): 4031, 4028, 4947, 4948, 6237, 6236, 6538, 6541, 6498, 6497, IHV116260, IHV116292, 540, 543, 544, T-71726, T-71727, T-71731, and T-71732. F=failed; M=missing; T=truncation; D=duplicated; NA=not available.

Ref. locus tag	GeP	BIGSdb GC	SeqSphere+	Product
BN867_00630	1,2,2,2,2,2,2,2,2,2,1,2,2,2,2,2,2,2,2	1,2,2,2,2,2,2,2,2,2,1,2,2,2,2,2,2,2,2	2,1,1,1,1,1,1,1,1,2,1,1,1,1,1,1,1	Hemerythrin-like iron-binding protein
BN867_00640	1,2,2,2,1,1,1,1,1,1,3,1,4,4,4,1,1,1,1	1,2,2,2,2,1,1,1,1,1,1,2,1,3,3,3,1,1,1,1	M,2,2,2,M,M,M,M,M,M,M,M,M,1,1,1,3,3,3,3	Hemerythrin-like iron-binding protein
BN867_00650	1,M,M,M,T,T,T,T,T,T,1,2,T,T,T,T,T,T,T	1,2,T,T,1,1,1,1,1,1,1,3,T,T,T,T,T,T,T	1,M,M,M,1,1,1,1,1,1,1,1,F,M,M,M,M,M,M,M	Hemerythrin-like iron-binding protein
BN867_01360	1,T,T,T,D,T,D,D,T,D,1,1,T,T,T,T,T,T,T	1,2,T,T,T,T,T,T,T,T,1,1,T,T,T,T,T,T,T	1,M,M,M,M,F,M,M,F,M,1,1,M,M,M,M,M,M,M	Methyl-accepting chemotaxis signal transduction protein
BN867_02370	1,M,M,M,D,D,D,T,D,D,1,2,T,T,T,M,M,M,M	1,2,T,T,T,T,T,T,T,T,1,3,T,T,T,T,T,T,T	1,M,M,M,F,F,F,F,F,F,1,2,M,M,M,M,M,M,M	Methyl-accepting chemotaxis signal transduction protein
BN867_05110	1,1,1,1,1,1,D,1,1,1,1,1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1	NA	Putative periplasmic protein
BN867_05120	1,1,1,1,1,1,D,1,1,1,1,1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1	NA	FIG00469903: hypothetical protein
BN867_05130	1,T,T,T,1,1,D,1,1,1,1,1,1,1,1,1,1,1,1	1,2,T,T,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1	NA	Filamentous haemagglutinin domain protein
BN867_05140	1,1,1,1,1,1,D,1,1,1,1,M,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1,1,2,1,1,1,1,1,1,1	NA	hypothetical protein
BN867_05150	1,2,2,2,1,1,D,1,T,T,1,1,1,1,1,3,3,3,3	1,2,2,2,1,1,1,1,1,1,1,1,1,1,1,3,3,3,3	NA	Putative hemolysin activation/secretion protein
BN867_06930	1,2,T,T,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1	1,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1	2,1,1,1,2,2,2,2,2,2,2,2,M,M,M,M,M,M,M	DNA adenine methylase
BN867_06950	1,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1	1,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1	1,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1	hypothetical protein
BN867_06960	1,2,2,2,1,1,1,1,3,3,1,4,1,D,1,4,4,4,4	1,2,2,2,1,1,1,1,3,3,1,4,1,T,1,4,4,4,4	1,M,M,M,1,1,1,1,F,F,1,2,M,M,M,2,2,2,2	phage repressor protein, putative
BN867_06990	1,2,2,2,1,1,1,1,1,1,1,1,1,T,1,1,1,1,1	1,2,2,2,1,1,1,1,1,1,1,1,1,T,1,1,1,1,1	1,2,2,2,1,1,1,1,1,1,1,1,1,M,1,1,1,1,1	FIG00471770: hypothetical protein
BN867_07950	1,T,T,T,T,T,D,T,N,N,1,1,1,1,1,1,1,1,1	1,2,T,T,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1	M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M	FIG00470444: hypothetical protein
BN867_08150	1,T,T,T,T,T,T,T,N,N,1,1,1,1,1,1,1,1,1	1,2,T,T,1,1,1,1,3,3,1,1,1,1,1,1,1,1,1	M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M	FIG00470444: hypothetical protein
BN867_09580	1,1,1,1,D,1,1,1,1,1,1,1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1	NA	Putative periplasmic protein
BN867_09590	1,1,1,1,D,1,1,1,1,1,1,1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1	NA	FIG00469903: hypothetical protein
BN867_09600	1,T,T,T,D,1,1,1,1,1,1,1,1,1,1,1,1,1,1	1,2,T,T,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1	NA	Filamentous haemagglutinin domain protein
BN867_09610	1,1,1,1,D,1,1,1,1,1,1,M,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1,1,2,1,1,1,1,1,1,1	NA	hypothetical protein
BN867_09620	1,2,2,2,D,1,1,1,T,T,1,1,1,1,1,3,3,3,3	1,2,2,2,2,1,1,1,1,1,1,1,1,1,1,1,3,3,3,3	NA	Putative hemolysin activation/secretion protein
BN867_09630	1,M,M,M,D,D,D,D,D,D,1,2,T,T,T,T,T,T,T	1,2,T,T,3,3,3,3,1,1,1,4,T,T,T,T,T,T,T	M,M,M,M,M,M,1,1,M,M,M,M,M,M,M,M,M,M,M	Hemerythrin-like iron-binding protein
BN867_09640	1,D,D,D,D,D,D,D,D,D,2,1,3,3,3,1,1,1,1	1,2,2,2,2,1,1,1,1,1,1,2,1,3,3,3,1,1,1,1	M,2,2,2,M,M,M,M,M,M,M,M,1,1,1,3,3,3,3	Hemerythrin-like iron-binding protein
BN867_09650	1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1	Hemerythrin-like iron-binding protein
BN867_13090	1,M,M,M,T,T,T,T,T,T,1,1,1,1,1,1,1,1,1	1,M,M,M,T,T,T,T,T,T,1,1,1,1,1,1,1,1,1	1,M,M,M,M,M,M,M,M,M,1,1,1,1,1,1,1,1,1	FIG00471635: hypothetical protein
BN867_13100	1,M,M,M,T,T,1,T,T,T,1,1,1,1,1,1,1,1,1	1,M,M,M,T,T,1,T,T,T,1,1,1,1,1,1,1,1,1	1,M,M,M,M,M,1,M,M,M,1,1,1,1,1,1,1,1,1	FIG00471635: hypothetical protein
BN867_13180	1,T,T,T,M,T,T,T,T,M,1,2,2,M,2,M,M,T,M	1,2,M,M,M,T,T,T,T,M,1,3,3,M,3,M,M,T,M	2,M,M,M,M,M,M,M,M,M,2,1,1,M,1,M,M,M,M	Motility accessory factor
BN867_13200	1,M,M,M,M,M,M,T,T,M,2,3,3,M,3,M,M,T,M	1,2,T,T,M,M,M,T,T,M,3,4,T,M,T,M,M,T,M	1,M,M,M,M,M,M,M,M,M,2,3,M,M,M,M,M,M,M	Motility accessory factor
BN867_13220	1,T,T,T,T,T,T,T,T,T,2,3,1,M,1,4,5,T,4	1,T,T,T,T,T,T,T,T,T,2,3,1,M,1,4,5,T,4	M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M	Flagellin
BN867_13230	1,T,T,T,T,T,T,T,T,T,2,T,3,M,3,1,4,T,1	1,T,T,T,T,T,T,T,T,T,2,3,3,M,3,1,2,M,1	M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M,M	Flagellin
BN867_13240	1,M,T,T,T,T,T,D,T,2,1,3,1,D,1,T,1,D,T	1,M,T,T,T,T,T,T,T,2,1,3,1,T,1,T,1,T,T	1,M,M,M,M,M,M,M,M,3,1,4,1,M,1,M,1,M,M	Motility accessory factor
BN867_13250	1,M,T,T,T,T,M,D,1,M,1,M,1,D,1,T,1,D,T	1,M,T,T,T,T,M,T,1,2,1,3,1,T,1,T,1,T,T	1,M,M,M,M,M,M,M,1,M,1,M,1,M,1,M,1,M,M	Motility accessory factor
BN867_14900	1,M,M,M,1,1,1,1,1,1,1,1,1,1,1,M,M,M,M	1,2,2,2,1,1,1,1,1,1,1,1,1,1,1,3,3,3,3	1,M,M,M,1,1,1,1,1,1,1,1,1,1,1,1,M,M,M,M	hypothetical protein
BN867_15290	1,M,M,M,T,M,T,T,M,T,1,2,T,M,T,T,T,T,T	1,M,M,M,T,T,T,T,T,T,1,2,T,M,T,T,T,T,T	1,M,M,M,M,M,M,F,M,M,1,2,M,M,M,M,M,M,M	Methyl-accepting chemotaxis signal transduction protein

FIGURES

Figure S1

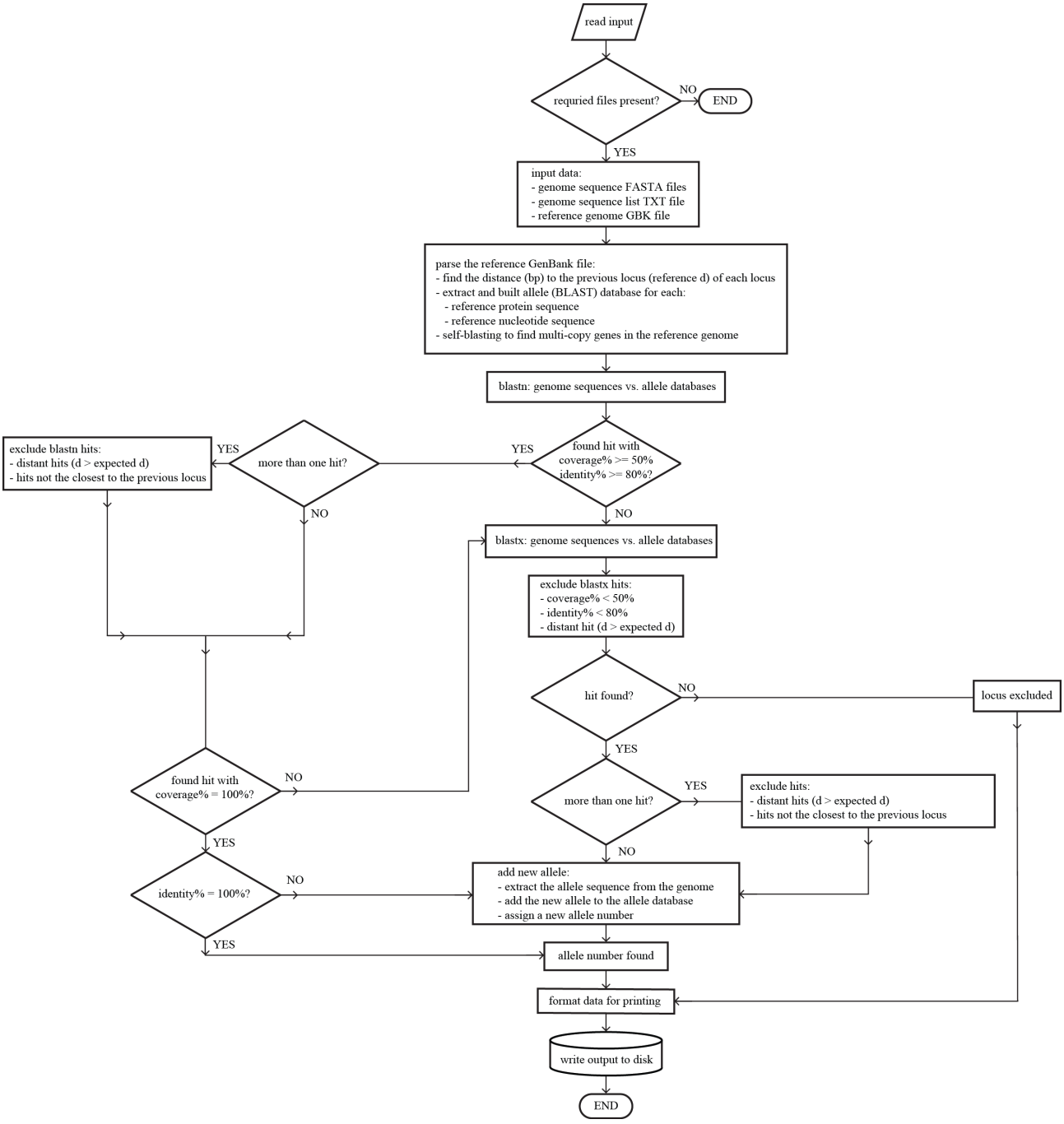

Figure S2

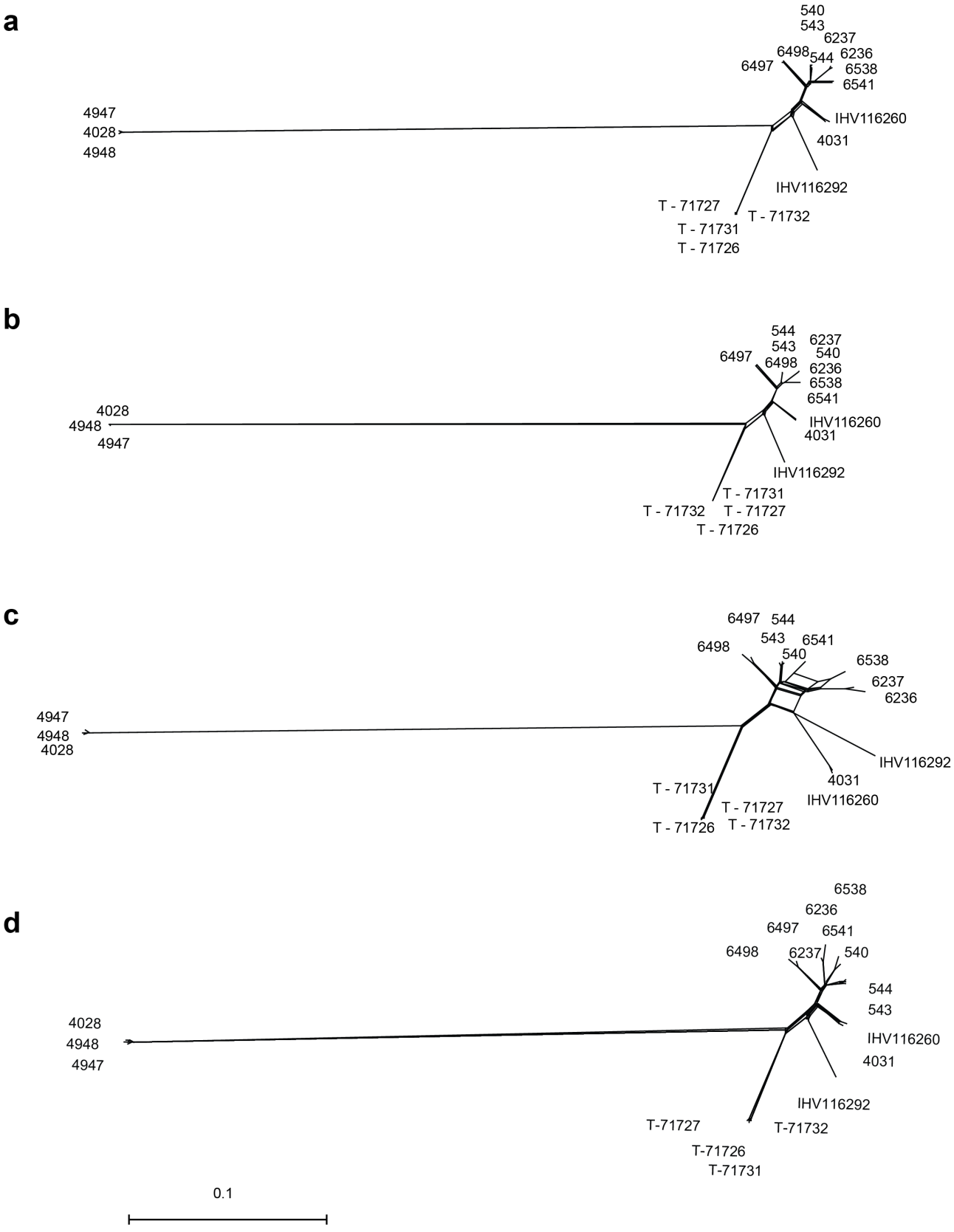
a
540
543
6237
6498
544
6236
6497
6538
6541
4947
4028
4948
IHV116260
4031
IHV116292
T - 71727
T - 71732
T - 71731
T - 71726
b
544
543
6237
6497
540
6498
6236
6538
6541
4028
4948
4947
IHV116260
4031
IHV116292
T - 71731
T - 71732
T - 71727
T - 71726
c
6497
544
543
6541
6498
540
6538
6237
6236
4947
4948
4028
IHV116292
4031
T - 71731
IHV116260
T - 71727
T - 71726
T - 71732
d
6538
6236
6497
6541
6498
6237
540
544
543
4028
4948
4947
IHV116260
4031
IHV116292
T-71727
T-71732
T-71726
T-71731
0.1

FIGURE CAPTURES

Figure S1. Flow chart of GeP logic. Standard symbols for constructing flow charts were used.

Figure S2. Split decomposition of the allelic profile of the 19 *C. jejuni* genomes generated by GeP (panel a), SeqSphere+ (panel b) and BIGSdb GC (panel c). The last graph (panel d) is also generated by BIGSdb GC, but with all the Error type II eliminated from the result.

REFERENCES

1. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.** 2009. BLAST+: architecture and applications. BMC bioinformatics **10:**421.
2. **Katoh K, Standley DM.** 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol **30:**772-780.
3. **Huson DH, Bryant D.** 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol **23:**254-267.
4. **Huson DH.** 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics **14:**68-73.
5. **Pritchard JK, Stephens M, Donnelly P.** 2000. Inference of population structure using multilocus genotype data. Genetics **155:**945-959.
6. **Corander J, Waldmann P, Marttinen P, Sillanpaa MJ.** 2004. BAPS 2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics **20:**2363-2369.
7. **Hänninen ML, Haajanen H, Pummi T, Wermundsen K, Katila ML, Sarkkinen H, Miettinen I, Rautelin H.** 2003. Detection and typing of *Campylobacter jejuni* and *Campylobacter coli* and analysis of indicator organisms in three waterborne outbreaks in Finland. Appl Environ Microbiol **69:**1391-1396.
8. **Revez J, Llarena AK, Schott T, Kuusi M, Hakkinen M, Kivisto R, Hänninen ML, Rossi M.** 2014. Genome analysis of *Campylobacter jejuni* strains isolated from a waterborne outbreak. BMC genomics **15:**768.
9. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA.** 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol **19:**455-477.
10. **Jolley KA, Maiden MC.** 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC bioinformatics **11:**595.
11. **Kohl TA, Diel R, Harmsen D, Rothganger J, Walter KM, Merker M, Weniger T, Niemann S.** 2014. Whole-genome-based *Mycobacterium tuberculosis* surveillance: a standardized, portable, and expandable approach. J Clin Microbiol **52:**2479-2486.
12. **Revez J, Zhang J, Schott T, Kivisto R, Rossi M, Hänninen ML.** 2014. Genomic variation between *Campylobacter jejuni* isolates associated with milk-borne-disease outbreaks. J. Clin. Microbiol. **52:**2782-2786.
13. **Jerome JP, Bell JA, Plovanich-Jones AE, Barrick JE, Brown CT, Mansfield LS.** 2011. Standing genetic variation in contingency loci drives the rapid adaptation of *Campylobacter jejuni* to a novel host. PloS one **6:**e16399.
14. **Revez J, Schott T, Llarena AK, Rossi M, Hänninen ML.** 2013. Genetic heterogeneity of *Campylobacter jejuni* NCTC 11168 upon human infection. Infect Genet Evol **16:**305-309.