2016 GMI Dry-lab PT Report

Contents

Introduction

This document summarizes the results of the dry-lab component of the 2016 Global Microbial Identifier (GMI) Proficiency Test (PT). For additional information about GMI and the various working groups please visit http://www.globalmicrobialidentifier.org

The objective of the dry-lab component was to assess the differences among laboratories in the detection of variants (e.g., single nucleotide polymorphisms (SNPs)) from the analysis of whole genome sequence data. Participants were provided three datasets of fastq files from each of three taxononic groups (*Listeria monocytogenes*, *Klebsiella pneumoniae*, and *Campylobacter jejuni*) and asked to analyze them with the current protocol implemented in their lab for detecting variants. In addition to answering an online survey regarding the type of analysis the participant performed, the participant also submitted a fasta formatted matrix of variants and a newick formatted tree file.

This document describes the analysis of those three sources of data - the survey, fasta matrix, and newick tree file.

Summary and Key Findings

- **A total of 215 results files were submitted (fasta or newick tree) (Table 1).**
- **Not surprisingly, participants differed in how they quality filtered (Fig. 1) and the methods they used to analyze the datasets (Figs. 2 & 3, Table 2).**
- **The number of positions within the fasta matrices differed greatly (Table 3).**
- **Despite differences in the size of the matrices and, in some cases, relative differences among samples, the majority of participants created trees that contained the same clusters of isolates (Tables 4 - 6).**

Participation

Table 1. The fasta and tree results that were analyzed per participant. A value of NA indicates that either the file was not provided or was provided but not usable (reasons a file may not have been usable include too many samples in the file, too few samples in the file, a format that could not be confidently coerced to either fasta or newick). See Data Curation subsection of the Methods section below for more information:

Lab	CJ_Fasta	CJ_Tree	KP_Fasta	KP_Tree	LM_Fasta	LM_Tree
GMI100	1	1	1	1	1	1
GMI102	NA	NA	NA	NA	NA	1
GMI104	1	1	1	1	1	1
GMI105	1	1	1	1	1	1
GMI106	1	1	1	1	1	1
GMI107	1	1	1	1	1	1
GMI110	1	1	1	1	1	1
GMI111	NA	NA	NA	NA	NA	NA
GMI114	NA	NA	NA	NA	NA	NA
GMI115	1	1	NA	NA	1	1
GMI116	1	1	1	1	1	1
GMI117	1	1	NA	NA	NA	NA
GMI118	1	1	NA	NA	1	1
GMI122	NA	NA	NA	NA	NA	NA
GMI65	NA	NA	NA	NA	NA	NA
GMI66	1	1	1	1	1	1
GMI67	1	NA	1	NA	1	NA
GMI70	1	1	1	1	1	1
GMI71	1	1	NA	NA	1	1
GMI72	1	1	1	1	NA	1
GMI73	1	1	1	1	1	1
GMI74	NA	1	NA	1	NA	1
GMI75	1	1	1	1	1	1
GMI77	1	1	1	1	1	1
GMI79	NA	1	NA	1	NA	NA
GMI80	1	1	1	1	1	1
GMI81	NA	1	NA	1	NA	1
GMI82	1	1	1	1	1	1
GMI83	1	NA	1	NA	1	NA
GMI84	1	1	1	1	1	1
GMI85	1	NA	1	NA	1	NA
GMI88	1	1	1	1	1	1
GMI89	NA	NA	NA	NA	NA	1
GMI90	NA	NA	1	1	1	1
GMI92	1	1	1	1	1	1
GMI93	1	1	1	1	1	1
GMI94	NA	NA	NA	NA	NA	NA
GMI95	1	1	NA	NA	NA	NA
GMI96	1	1	1	1	1	1
GMI97	NA	1	NA	1	NA	1
GMI98	1	1	1	1	1	1

Diversity of the Methods Being Used

Figures 1-3. Charts illustrating the diversity of methods and practices employed for detecting variant from WGS data.

Figure 1. Responses to question about filtering and detection methods

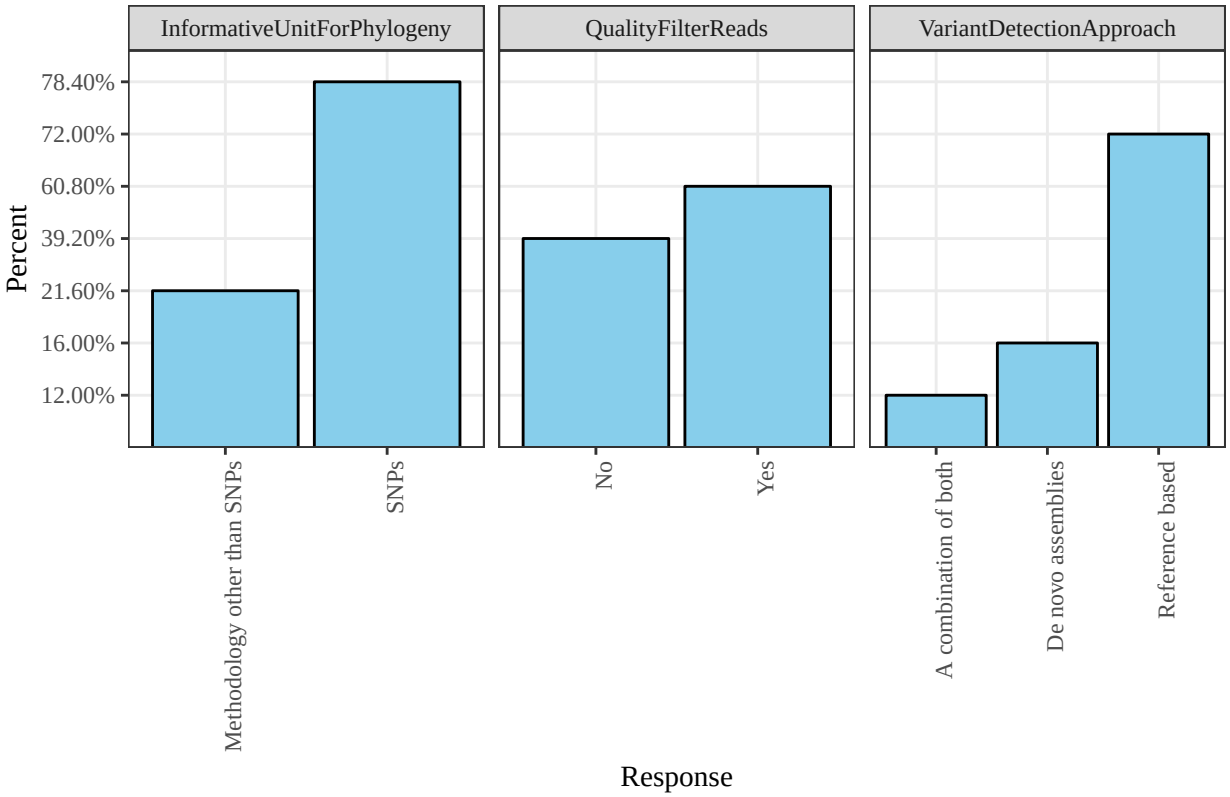


Figure 2. Optimality Criterion for Inferring Phylogeny

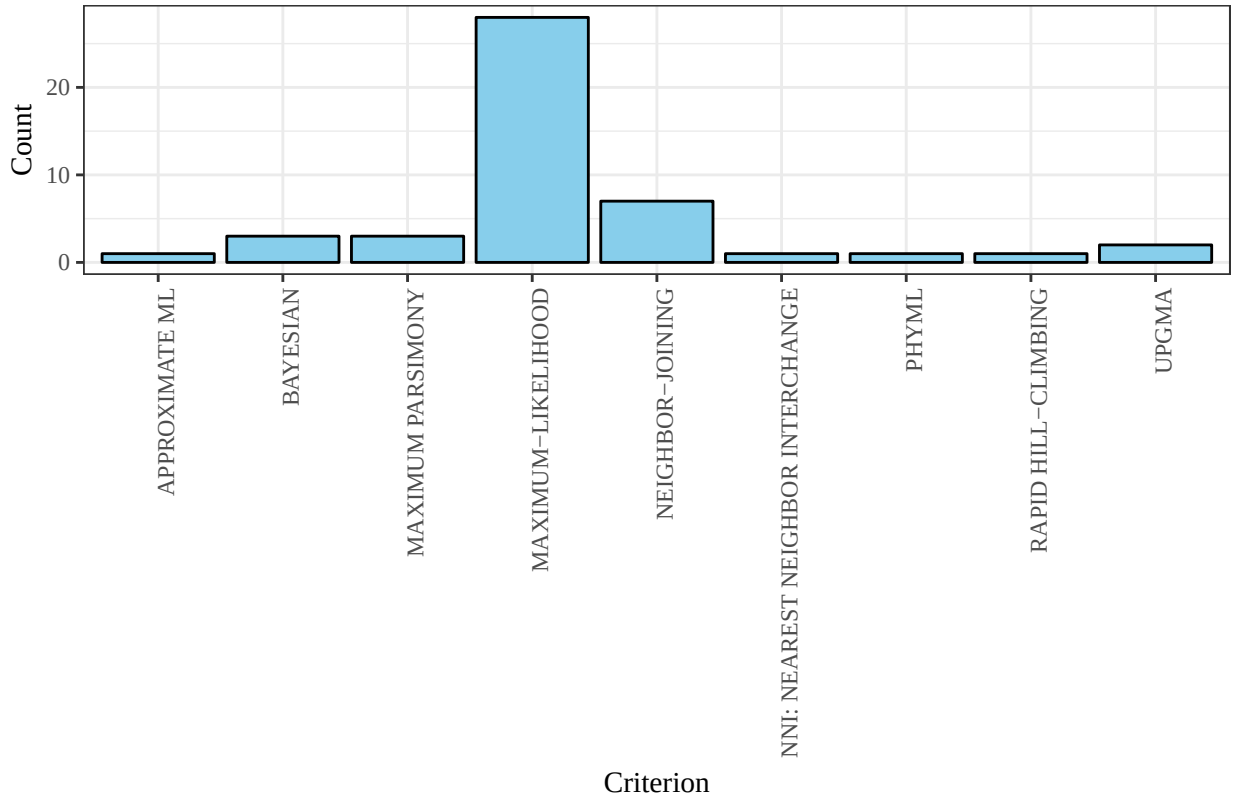


Figure 3. Software for Inferring Phylogeny

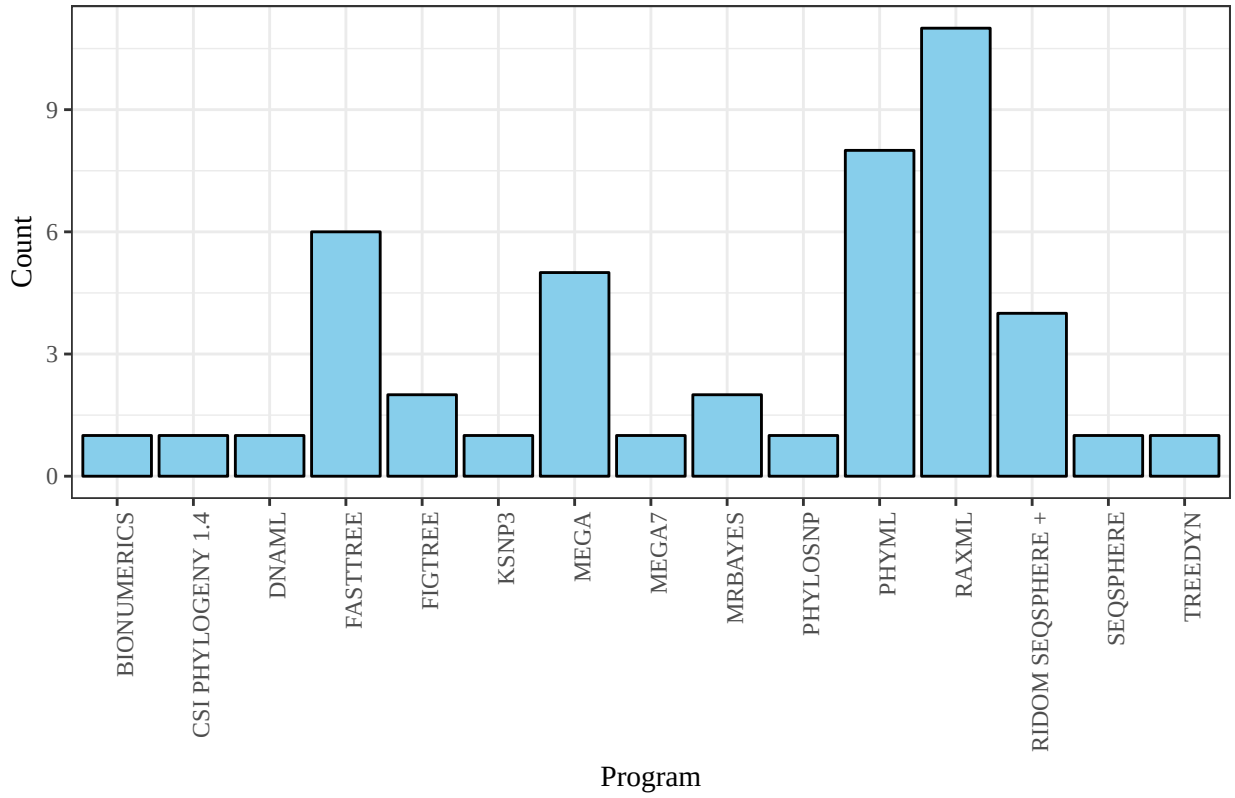


Table 2. Diversity of mapping and variant detection software

	Programs
1	Bbmap
2	Bionumerics
3	Bowtie2
4	Bowtie2, GATK
5	Bowtie2, VarScan
6	BWA
7	BWA-MEM FreeBayes, Galaxy
8	BWA, FreeBayes
9	BWA, GATK
10	BWA, Samtools, PicardTools
11	BWA, Varscan
12	CFSAN SNP pipeline v0.7.0
13	CLC assembly cell
14	CSIPhylogeny
15	Parsnp, Gubbins
16	SeqSphere, ridom
17	SMALT
18	SNIPPY, nullarbor
19	SNVPhyl

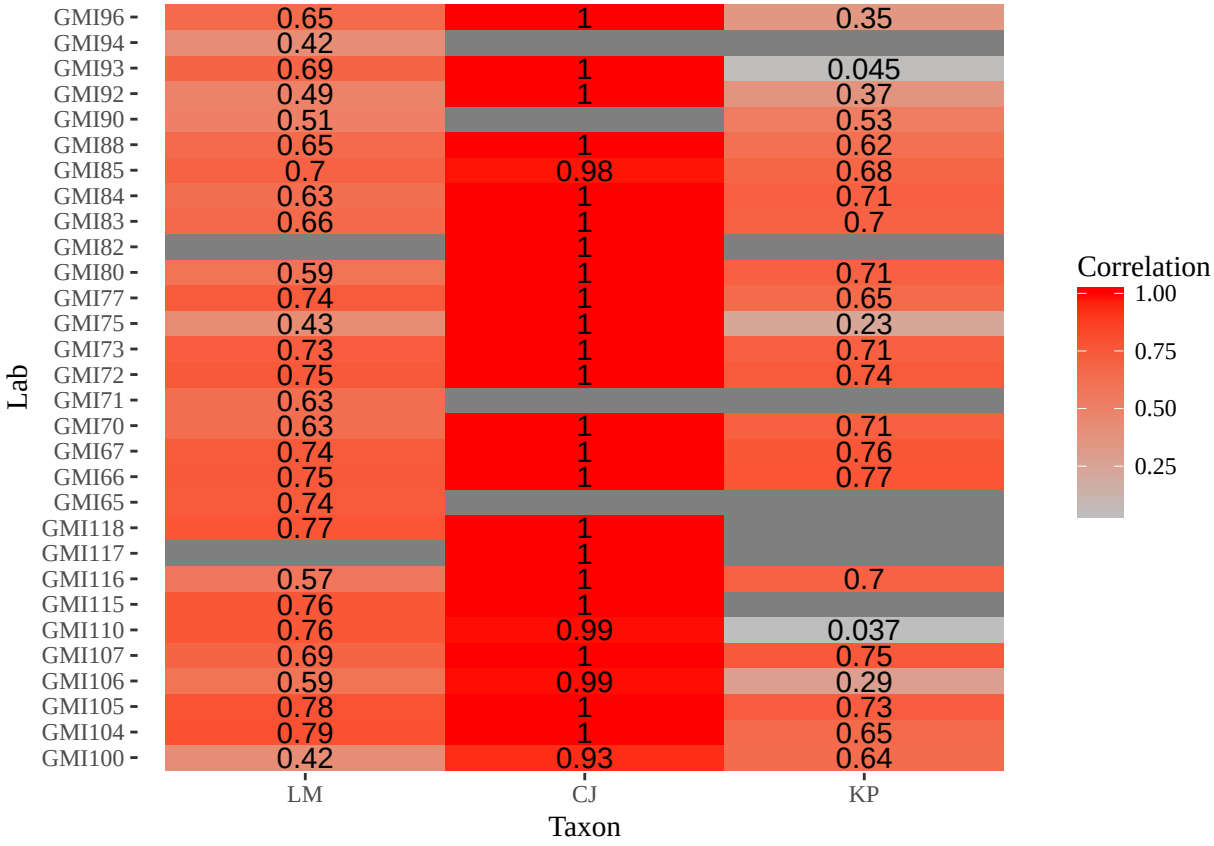
Size and Information Content of SNP Matrices

The SNP matrices differed greatly in the number of positions (Table 3, Fig. 2) and to a lesser extent the correlation between them in the pairwise SNP differences among samples (Fig. 3). The former is perhaps not surprising given the diversity of methods being employed (Fig. 1), how different references may have been used, and the various filtering of SNPs (e.g., on SNP density) that labs may be employing. The fact that there is a good correlation between labs in the pairwise differences, suggests that different size matrices still contain the same relative information as to how similar samples are (e.g., in one lab two pairs of samples may differ by 10 and 50 SNPs but in another lab they may only differ by 1 and 5 SNPs).

Table 3. Table of the number of positions in each SNP matrix (CJ = *Campylobacter jejuni*; lm = *Klebsiella pneumoniae* ; LM = *Listeria monocytogenes*).

Lab	CJ	KP	LM
GMI100	1851	126	1652
GMI104	728	20	63
GMI105	55	80	71
GMI106	69	555	235
GMI107	1879	95	146
GMI110	39	1416	77
GMI115	1644	NA	107
GMI116	2084	178	1811
GMI117	747	NA	NA
GMI118	1268	NA	77
GMI66	1516	91	98
GMI67	1516	91	101
GMI70	1696	126	1616
GMI71	NA	NA	294
GMI72	478	87	NA
GMI73	205	87	77
GMI75	1492548	4626433	2566491
GMI77	728	20	63
GMI80	1734	126	144
GMI82	1149	99	NA
GMI83	1619756	5581932	2941727
GMI84	1205	113	699
GMI85	43	97	85
GMI88	1260	167	233
GMI90	NA	105	1422
GMI92	1680	120	957
GMI93	728	253	63
GMI95	648	NA	NA
GMI96	728	20	63
GMI98	1620929	5582195	2941547

Figure 2. Heatmap of the average correlation between labs in the pairwise distance between samples for each of the three datasets. An empty cell denotes that lab either did not provide a SNP matrix for that taxon or the one provided could not be analyzed.



Results of Cluster Detection Analyses

On average 80% of Labs produced topologies within which specific clusters of individuals could be found; congruence among the labs in the clustering of *Klebsiella pneumoniae* was particularly low and reflects the high similarity of all isolates in that database. Figures 4 - 6 show the results of each lab for each taxonomic group. This suggests that despite the diversity of methods being employed and differences in the size of matrices, the clustering of individuals is similar among labs.

Table 4. Results of tests to determine whether each *Campylobacter jejuni* topology clustered specific samples together (100% correctly clustered cluster1). A value of TRUE means the individuals were correctly clustered; FALSE means the cluster containing members of a defined cluster also included those that did not belong to it.

Lab	Cluster1
GMI100_CJ	TRUE
GMI104_CJ	TRUE
GMI105_CJ	TRUE
GMI106_CJ	TRUE
GMI107_CJ	TRUE
GMI110_CJ	TRUE
GMI115_CJ	TRUE
GMI116_CJ	TRUE
GMI117_CJ	TRUE
GMI118_CJ	TRUE
GMI66_CJ	TRUE
GMI70_CJ	TRUE
GMI71_CJ	TRUE
GMI72_CJ	TRUE
GMI73_CJ	TRUE
GMI74_CJ	TRUE
GMI75_CJ	TRUE
GMI77_CJ	TRUE
GMI79_CJ	TRUE
GMI80_CJ	TRUE
GMI81_CJ	TRUE
GMI82_CJ	TRUE
GMI84_CJ	TRUE
GMI88_CJ	TRUE
GMI92_CJ	TRUE
GMI93_CJ	TRUE
GMI95_CJ	TRUE
GMI96_CJ	TRUE
GMI97_CJ	TRUE
GMI98_CJ	TRUE

Table 5. Results of tests to determine whether each *Klebsiella pneumoniae* topology clustered specific samples together (73% of Labs correctly clustered cluster1; 57% correctly clustered cluster2). A value of TRUE means the individuals were correctly clustered; FALSE means the cluster containing members of a defined cluster also included those that did not belong to it.

Lab	Cluster1	Cluster2
GMI100_KP	TRUE	TRUE
GMI104_KP	TRUE	FALSE
GMI105_KP	TRUE	TRUE
GMI106_KP	FALSE	FALSE
GMI107_KP	TRUE	TRUE
GMI110_KP	FALSE	FALSE
GMI116_KP	TRUE	TRUE
GMI66_KP	TRUE	TRUE
GMI70_KP	TRUE	TRUE
GMI72_KP	TRUE	TRUE
GMI73_KP	TRUE	TRUE
GMI74_KP	FALSE	FALSE
GMI75_KP	FALSE	FALSE
GMI77_KP	TRUE	FALSE
GMI79_KP	FALSE	FALSE
GMI80_KP	TRUE	TRUE
GMI81_KP	TRUE	FALSE
GMI82_KP	TRUE	TRUE
GMI84_KP	TRUE	TRUE
GMI88_KP	TRUE	TRUE
GMI90_KP	TRUE	TRUE
GMI92_KP	TRUE	TRUE
GMI93_KP	TRUE	TRUE
GMI96_KP	TRUE	FALSE
GMI97_KP	FALSE	FALSE
GMI98_KP	FALSE	FALSE

Table 6. Results of tests to determine whether each *Listeria monocytogenes* topology clustered specific samples together (80% of Labs correctly clustered cluster1; 93% correctly clustered cluster2). A value of TRUE means the individuals were correctly clustered; FALSE means the cluster containing members of a defined cluster also included those that did not belong to it.

Lab	Cluster1	Cluster2
GMI100_LM	TRUE	TRUE
GMI102_LM	FALSE	TRUE
GMI104_LM	TRUE	TRUE
GMI105_LM	TRUE	TRUE
GMI106_LM	TRUE	TRUE
GMI107_LM	TRUE	TRUE
GMI110_LM	TRUE	FALSE
GMI115_LM	TRUE	TRUE
GMI116_LM	TRUE	TRUE
GMI118_LM	TRUE	TRUE
GMI66_LM	TRUE	TRUE
GMI70_LM	TRUE	TRUE
GMI71_LM	TRUE	TRUE
GMI72_LM	TRUE	TRUE
GMI73_LM	TRUE	TRUE
GMI74_LM	FALSE	TRUE
GMI75_LM	FALSE	TRUE
GMI77_LM	TRUE	TRUE
GMI80_LM	TRUE	TRUE
GMI81_LM	FALSE	TRUE
GMI82_LM	TRUE	TRUE
GMI84_LM	TRUE	TRUE
GMI88_LM	TRUE	TRUE
GMI89_LM	TRUE	TRUE
GMI90_LM	FALSE	TRUE
GMI92_LM	TRUE	TRUE
GMI93_LM	TRUE	TRUE
GMI96_LM	TRUE	TRUE
GMI97_LM	TRUE	FALSE
GMI98_LM	FALSE	TRUE

Conclusions

The results from the dry-lab component of the 2016 GMI PT highlight the diversity of bioinformatic tools that are being employed around the world to analyze whole-genome sequence data of bacteria that are of importance to public health and food safety. Not surprisingly these methods do not produce the same data objects (variant positions and SNP matrices) from which phylogenetic trees (topologies) are inferred. However, despite those differences, the topologies submitted by the more than 40 participants in this PT clustered samples quite similarly (>80% of trees submitted by participants clustered samples correctly) suggesting that a labs would reach similar conclusions when the methods are applied to traceback and sourcetraking investigations.

These results suggest that based on internal validation studies, individual centres will be able to define sensible thresholds for determining clusters of isolates. However the fact that the absolute number of variants and branch lengths reported differ markedly between centres has implications for public health since thresholds may vary between labs. As the technology continues to be used, a standardised approach will likely emerge within which thresholds will be decided upon that will facilitate congruence among centre-specific pipelines in the conclusions that are reached.

Methods

Data Curation

There were many differences in terms of syntax of names and formats among the SNP matrices and trees submitted by the participants. A number of steps were taken to correct as many inconsistencies as possible but unfortunately some results that were submitted could not be analyzed. The most likely cause was that the number of samples in the SNP matrix and/or tree did not match the expected number. In future PTs we will be more explicit as to what should be included in the results files and how samples should be named.

SNP Distance Calculation

SNP differences among samples were calculated using the "N" model within the dist.dna function in the R (R Core Team, 2015) package ape (Paradis et al., 2004).

Cluster Detection Analysis

Within each dataset a number of clusters were defined that we then determined if they were present in each tree. The clusters were:

- Listeria monocytogenes cluster 1 = LM13, LM3, LM14, LM20, LM1, LM16
- Listeria monocytogenes cluster 2 = LM7, LM12, LM8
- Klebsiella pneumoniae cluster 1 = KP5, KP9
- Klebsiella pneumoniae cluster 2 = KP16, KP11
- Campylobacter jejuni cluster 1 = CJ1, CJ2

Each tree was then rooted on the same individual and the node that united all members of a cluster was determined using the getMRCA function in the ape package. The members of the clade defined by that node was then determined using the clade.members function in the R package casper (Orme et al., 2013). If those belonging to an a priori cluster differed from those found in the clade uniting all of them on a tree, a value of FALSE was returned indicating that the tree did not contain the correct cluster; otherwise a value of TRUE was returned.

Acknowledgements

The following members of Working Group 4 were integral in the development and deployment of the 2016 GMI PT: James Pettengill (Biostatistics and Bioinformatics Staff, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration), Oksana Lukjancenko (Technical University of Denmark, National Food Institute), Errol Strain (Biostatistics and Bioinformatics Staff, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration), Susanne Karlsmose Pedersen (Technical University of Denmark, National Food Institute), Rene S. Hendriksen (Technical University of Denmark, National Food Institute), Bill Wolfgang (Department of Health, Wadsworth Center, New York State); Henrik Torkil Westh (Department of Clinical Microbiology, Amager and Hvidovre Hospital, Denmark); Vitali Sintchenko (Sydney Medical School, Australia); Jacob Moran-Gilad (Ministry of Health, Israel); William Hsiao (Department of Pathology and Laboratory Medicine, Vancouver, BC Canada); Brian Beck (Microbiologics, Inc.); Eija Trees (US CDC, Atlanta, GA); Isabel Cuesta de la plaza (Instituto de Salud Carlos III, Spain); Angel Zaballos (Instituto de Salud Carlos III, Spain); Jorge De La Barrera Martinez (Instituto de Salud Carlos III, Spain).

Referencing This Document

This document will be placed on the GMI website where the corresponding url can be used to reference it.

Contacts for questions regarding 2016 GMI PT Dry-lab Report

Questions regarding this document should be directed to James Pettengill (james.pettengill@fda.hhs.gov) or Susanne Karlsmose Pedersen (suska@food.dtu.dk)

Citations

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Paradis E., Claude J. & Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289-290.

David Orme, Rob Freckleton, Gavin Thomas, Thomas Petzoldt, Susanne Fritz, Nick Isaac and Will Pearse (2013). caper: Comparative Analyses of Phylogenetics and Evolution in R. R package version 0.5.2. http://CRAN.R-project.org/package=caper