2015 GMI PT Dry-lab Analyses and Report

James B Pettengill[1], Anthony Underwood[2], Oksana Lukjancenko[3], Errol Strain[1], Susanne Karlsmose Pedersen[3], Rene S. Hendriksen[3]

[1]Biostatistics and Bioinformatics Staff, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration

[2]Bioinformatics Unit, Infectious Disease Informatics, National Infection Service, Public Health England

[3]Technical University of Denmark, National Food Institute, WHO Collaborating Center for Antimicrobial Resistance in Food borne Pathogens and European Union Reference Laboratory for Antimicrobial Resistance, Kgs. Lyngby, Denmark

Contents

Introduction

This document summarizes the results of the dry-lab component of the 2015 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 and asked to analyze them with the current protocol implemented in their lab for detecting such 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 source of data - the survey, fasta matrix, and newick tree file.

Summary and Key Findings

- A total of 190 results files were submitted with a relatively even distribution across the three taxonomic groups and file types (fasta or newick tree) (Table 1).
- Not surprisingly, there are a diversity of algorithms being employed to, for example, map reads and infer a phylogeny. Participants also differed in the choices they made with respect to quality filtering and contamination checking (Figure 1).
- Within a given taxonomic group the number of positions within the fasta matrices differed greatly (Table 2).
- However, the matrices carry similar information content in terms of the relative magnitude of differences between samples (Fig. 2)
- Despite differences in the size of the matrices and, in some cases, relative differences among samples, the majority of participants created trees that did contain the clusters we were interested in detecting (Tables 3 - 5).
- Details on the methods and analyses performed can be found at the end of this document.

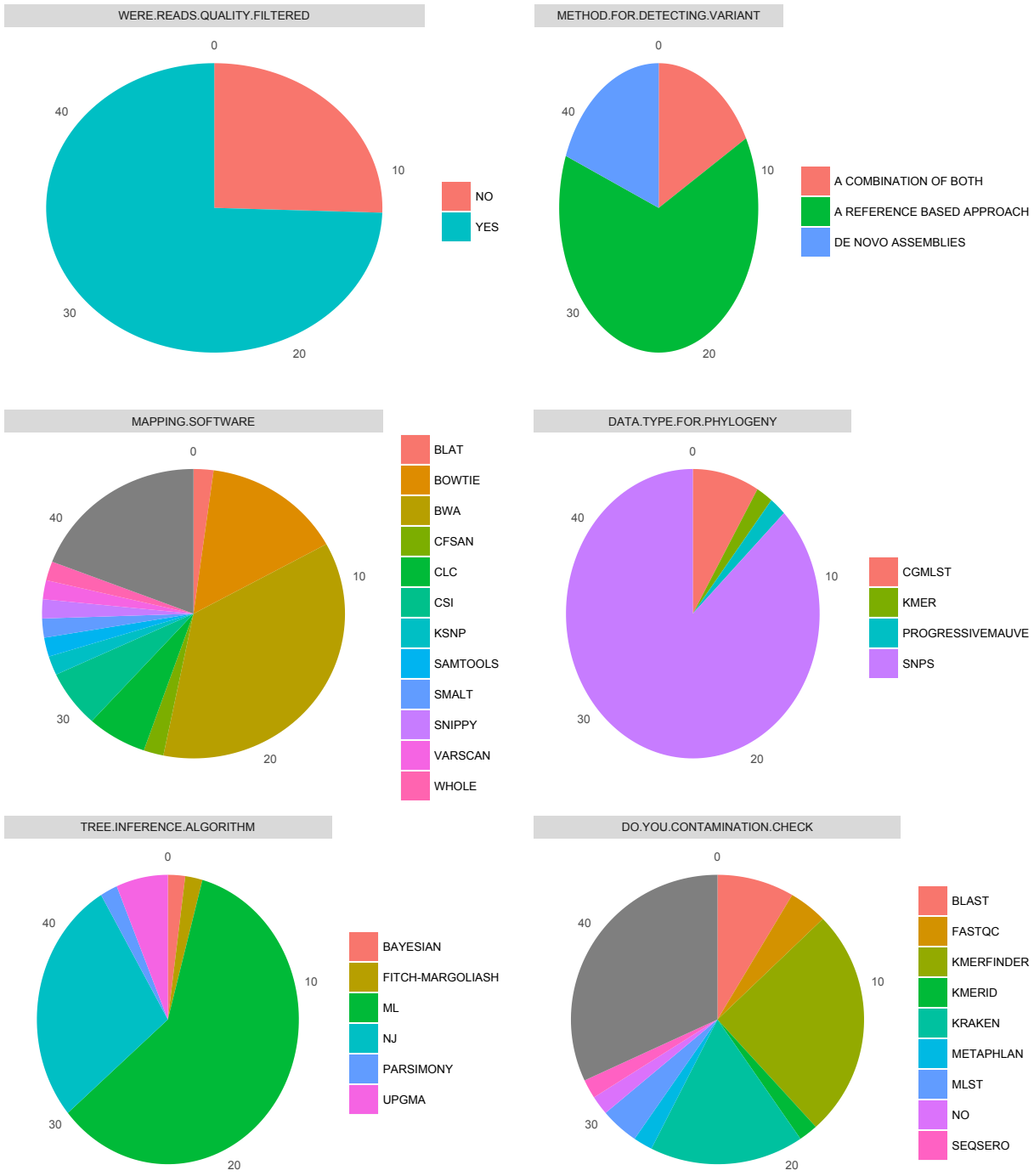
Participation

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 coerced to either fasta or newick):

LAB	EC.FASTA	EC.TREE	SA.FASTA	SA.TREE	ST.FASTA	ST.TREE	TOTAL
GMI02	1	1	1	1	1	1	6
GMI04	1	1	1	1	1	1	6
GMI06	1	NA	1	1	1	1	5
GMI10	1	1	1	1	1	1	6
GMI14	1	NA	NA	NA	1	1	3
GMI17	NA	1	NA	1	NA	1	3
GMI26	1	1	1	1	1	1	6
GMI34	NA	1	NA	1	NA	1	3
GMI39	1	1	1	1	1	1	6
GMI42	1	1	NA	NA	1	NA	3
GMI43	1	1	1	1	1	1	6
GMI46	1	NA	1	NA	1	1	4
GMI48	1	1	1	1	1	1	6
GMI58	1	1	NA	NA	1	NA	3
GMI59	1	1	1	1	1	1	6
GMI13	1	1	1	1	1	1	6
GMI15	NA	1	NA	1	NA	1	3
GMI16	1	1	1	1	1	1	6
GMI21	1	1	1	1	1	1	6
GMI22	1	1	1	1	1	1	6
GMI24	NA	1	NA	1	NA	1	3
GMI27	1	1	1	1	1	1	6
GMI28	NA	NA	NA	NA	NA	1	1
GMI30	1	1	1	1	1	1	6
GMI31	NA	NA	NA	NA	1	1	2
GMI32	1	1	1	1	1	1	6
GMI33	1	1	1	1	1	1	6
GMI35	1	1	1	1	NA	NA	4
GMI37	1	NA	1	1	1	1	5
GMI38	NA	1	NA	1	NA	1	3
GMI40	1	1	1	1	1	1	6
GMI44	1	1	1	1	1	1	6
GMI45	1	1	1	1	1	1	6
GMI47	1	1	1	1	1	1	6
GMI50	1	1	NA	NA	1	NA	3
GMI51	1	1	NA	NA	1	NA	3
GMI55	NA	NA	1	NA	1	1	3
GMI61	NA	NA	NA	NA	1	NA	1
GMI63	NA	NA	1	NA	1	1	3
GMI7	1	1	1	1	1	1	6
GMI8	1	1	1	1	1	1	6
GRAND TOTAL	31	32	28	30	34	35	190

Diversity of the Methods Being Used

Figure 1. Pie charts illustrating the diversity of methods and practices employed for detecting variant from WGS data.



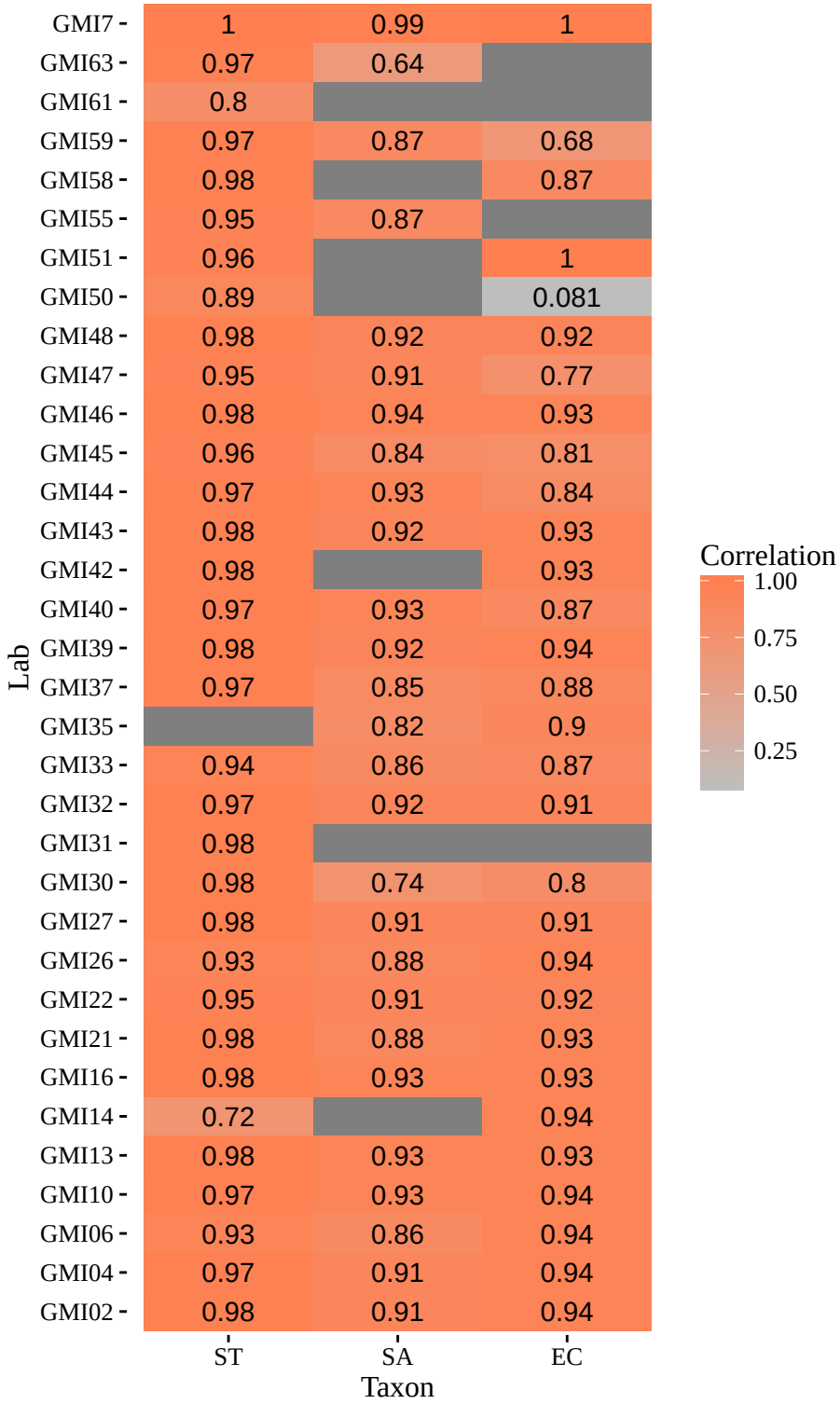
Size and Information Content of SNP Matrices

The SNP matrices differed greatly in the number of positions (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 2. Table of the number of positions in each SNP matrix (EC = E. coli; SA = S. aureus; ST = S. typhimurium). (NB: The SNP matrices from GMI21 and GMI45 are most likely all position matrices rather than just variant positions)

Lab	EC	SA	ST
GMI02	25731	1383	8968
GMI04	25731	1383	8968
GMI06	43264	6226	5822
GMI10	13083	1797	12902
GMI14	14687	NA	1431
GMI26	92831	6164	31044
GMI39	52590	2672	16034
GMI42	9460	NA	12884
GMI43	38532	4163	16562
GMI46	63273	2341	9958
GMI48	67034	2063	14080
GMI58	79231	NA	19656
GMI59	23561	2715	14199
GMI13	9276	1628	8746
GMI16	55473	2122	13630
GMI21	5187829	2837196	5090636
GMI22	33416	1597	13066
GMI27	33664	2130	13297
GMI30	607217	11881	12733
GMI31	NA	NA	4141
GMI32	14667	25949	28164
GMI33	71822	5420	21668
GMI35	6706	1334	NA
GMI37	73355	2897	14294
GMI40	45725	2033	11180
GMI44	35039	1836	9446
GMI45	5183821	2836332	5088344
GMI47	20707	1805	12198
GMI50	84	NA	1300
GMI51	35521	NA	10042
GMI55	NA	1644	9102
GMI61	NA	NA	24
GMI63	NA	2834703	5077509
GMI7	21731	1673	9192
GMI8	15972	1851	12979

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 93% of Labs produced topologies within which specific clusters of individuals could be found. 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 3. Results of tests to determine whether each E. coli topology clustered specific samples together (97% correctly clustered cluster1; 91% 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
GMI02	TRUE	TRUE
GMI04	TRUE	TRUE
GMI10	TRUE	TRUE
GMI17	FALSE	FALSE
GMI26	TRUE	TRUE
GMI34	TRUE	TRUE
GMI39	TRUE	TRUE
GMI42	TRUE	TRUE
GMI43	TRUE	TRUE
GMI48	TRUE	TRUE
GMI58	TRUE	TRUE
GMI59	TRUE	TRUE
GMI13	TRUE	TRUE
GMI15	TRUE	FALSE
GMI16	TRUE	TRUE
GMI21	TRUE	TRUE
GMI22	TRUE	TRUE
GMI24	TRUE	TRUE
GMI27	TRUE	TRUE
GMI30	TRUE	FALSE
GMI32	TRUE	TRUE
GMI33	TRUE	TRUE
GMI35	TRUE	TRUE
GMI38	TRUE	TRUE
GMI40	TRUE	TRUE
GMI44	TRUE	TRUE
GMI45	TRUE	TRUE
GMI47	TRUE	TRUE
GMI50	TRUE	TRUE
GMI51	TRUE	TRUE
GMI7	TRUE	TRUE
GMI8	TRUE	TRUE

Table 4. Results of tests to determine whether each S. aureus topology clustered specific samples together (93% of Labs correctly clustered cluster1; 93% correctly clustered cluster2; 97% correctly clustered cluster3). 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	Cluster3
GMI02	TRUE	TRUE	TRUE
GMI04	TRUE	TRUE	TRUE
GMI06	TRUE	TRUE	TRUE
GMI10	TRUE	TRUE	TRUE
GMI17	FALSE	FALSE	FALSE
GMI26	TRUE	TRUE	TRUE
GMI34	TRUE	TRUE	TRUE
GMI39	TRUE	TRUE	TRUE
GMI43	TRUE	TRUE	TRUE
GMI48	TRUE	TRUE	TRUE
GMI59	TRUE	TRUE	TRUE
GMI13	TRUE	TRUE	TRUE
GMI15	TRUE	TRUE	TRUE
GMI16	TRUE	TRUE	TRUE
GMI21	TRUE	TRUE	TRUE
GMI22	TRUE	TRUE	TRUE
GMI24	TRUE	TRUE	TRUE
GMI27	TRUE	TRUE	TRUE
GMI30	TRUE	FALSE	TRUE
GMI32	TRUE	TRUE	TRUE
GMI33	TRUE	TRUE	TRUE
GMI35	FALSE	TRUE	TRUE
GMI37	TRUE	TRUE	TRUE
GMI38	TRUE	TRUE	TRUE
GMI40	TRUE	TRUE	TRUE
GMI44	TRUE	TRUE	TRUE
GMI45	TRUE	TRUE	TRUE
GMI47	TRUE	TRUE	TRUE
GMI7	TRUE	TRUE	TRUE
GMI8	TRUE	TRUE	TRUE

Table 5. Results of tests to determine whether each S. typhimurium topology clustered specific samples together (97% of Labs correctly clustered cluster1; 86% 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
GMI02	TRUE	FALSE
GMI04	TRUE	FALSE
GMI06	TRUE	TRUE
GMI10	TRUE	TRUE
GMI14	TRUE	TRUE
GMI17	FALSE	FALSE
GMI26	TRUE	TRUE
GMI34	TRUE	TRUE
GMI39	TRUE	TRUE
GMI43	TRUE	FALSE
GMI46	TRUE	TRUE
GMI48	TRUE	TRUE
GMI59	TRUE	TRUE
GMI13	TRUE	TRUE
GMI15	TRUE	TRUE
GMI16	TRUE	TRUE
GMI21	TRUE	TRUE
GMI22	TRUE	TRUE
GMI24	TRUE	TRUE
GMI27	TRUE	TRUE
GMI28	TRUE	TRUE
GMI30	TRUE	TRUE
GMI31	TRUE	TRUE
GMI32	TRUE	TRUE
GMI33	TRUE	TRUE
GMI37	TRUE	TRUE
GMI38	TRUE	TRUE
GMI40	TRUE	TRUE
GMI44	TRUE	TRUE
GMI45	TRUE	TRUE
GMI47	TRUE	TRUE
GMI55	TRUE	TRUE
GMI63	TRUE	FALSE
GMI7	TRUE	TRUE
GMI8	TRUE	TRUE

Conclusions

The results from the dry-lab component of the 2015 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 (>93% of participants clustered samples correctly) suggesting that a vast majority of 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. We took a number of steps to correct as many inconsistencies as we could 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:

- S. aureus cluster 1 = SAH582, SAH605, SAH604, SAH602, SAM1048 , SAH597, SAH600, SAH599, SAH587, SAH596, SAH570
- S. aureus cluster 2 = SAM767, SAM774, SAM775, SAM760
- S. aureus cluster 3 = SAM1313, SAM1353
- E. coli cluster 1 = EC002156, EC002151
- E. coli cluster 2 = EC002117, EC002118, EC002116
- S. typhimurium cluster 1 = ST000026, ST000024
- S. typhimurium cluster 2 = ST003354, ST003377

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 also integral in the development and deployment of the 2015 GMI PT: 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.

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