Downloaded from orbit.dtu.dk on: Apr 25, 2024

DTU Library

The proficiency test (pilot) report of the global microbial identifier (GMI) initiative, year 2014

Hendriksen, Rene S.; Karlsmose Pedersen, Susanne; Larsen, Mette Voldby; Neubert Pedersen, Jamie; Lukjancenko, Oksana; Kaas, Rolf Sommer; Leekitcharoenphon, Pimlapas; Bergmark, Lasse; Hansen, Inge Marianne; Sintchenko, Vitali

Total number of authors:
23

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Hendriksen, R. S., Karlsmose Pedersen, S., Larsen, M. V., Neubert Pedersen, J., Lukjancenko, O., Kaas, R. S., Leekitcharoenphon, P., Bergmark, L., Hansen, I. M., Sintchenko, V., Wolfgang, W. J., Westh, H. T., Moran-Gilad, J., Hsiao, W., Cuesta, I., Barrera, J., Zaballos, A., Olson, N. D., Beck, B., ... Pettengill, J. (2016). *The proficiency test (pilot) report of the global microbial identifier (GMI) initiative, year 2014*. National Food Institute, Technical University of Denmark.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

THE PROFICIENCY TEST (PILOT) REPORT OF THE GLOBAL MICROBIAL IDENTIFIER (GMI) INITIATIVE, YEAR 2014

DTU Food
National Food Institute

THE PROFICIENCY TEST (PILOT) REPORT OF THE GLOBAL MICROBIAL IDENTIFIER INITIATIVE YEAR 2014

Rene S. Hendriksen[1], Susanne Karlsmose Pedersen[1], Mette Voldby Larsen[1], Jamie Neubert Pedersen[1], Oksana Lukjancenko[1], Rolf Sommer Kaas[1], Pimlapas Leekitcharoenphon[1], Lasse Bergmark[1], Inge Marianne Hansen[1], Vitali Sintchenko[2,3] William J. Wolfgang[4], Henrik Torkil Westh[5], Jacob Moran-Gilad[6], William Hsiao[7], Isabel Cuesta[8], Jorge Barrera[8], Angel Zaballos[9], Nathanael David Olson[10], Brian Beck[11], Anthony Underwood[12], Frank M. Aarestrup[1], Errol Strain[13], James Pettengill[13] and on behalf of the Global Microbial Identifier initiative's Working Group 4 (GMI-WG4);

[1]Technical University of Denmark, National Food Institute, Research Group of Genomic Epidemiology, Kgs. Lyngby, Denmark

[2]Sydney Medical School and Marie Bashir Institute for Infectious Diseases and Biosecurity, University of Sydney, Sydney, Australia

[3]Centre for Infectious Diseases and Microbiology – Public Health, Institute of Clinical Pathology and Medical Research – Pathology West, Westmead Hospital, Sydney, Australia

[4]Bacteriology Laboratory, Wadsworth Center, New York State Department of Health, Albany, New York, USA

[5]Hvidovre Hospital, Department of Clinical Microbiology, Hvidovre, Denmark

[6]Public Health Services, Ministry of Health & Faculty of Health Sciences, Ben-Gurion University, Israel

[7]BCCDC Public Health Microbiology & Reference Laboratory Clinical Assistant Professor, Pathology & Laboratory Medicine, UBC, Vancouver, BC, Canada

[8]National Center for Microbiology, Unidad de Bioinformática, Institute of Health Carlos III Carretera Majadahonda, Madrid, Spain

[9]Centro Nacional de Microbiología-ISCIII, Unidad de Genómica, Carretera Majadahonda-Pozuelo, Majadahonda, Madrid, Spain

[10]Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, MD, USA

[11]Microbiologics, Inc. Saint Cloud, Minnesota, USA

[12]Public Health England, Infectious Disease Informatics, Microbiology Services, Colindale, London

[13]Biostatistics Branch, US Food and Drug Administration, College Park, Maryland, USA

Opinions expressed in this paper are the authors' and do not necessarily reflect the policies and views of NIST, or affiliated venues. Certain commercial equipment, instruments, or materials are identified in this paper only to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. Official contribution of NIST; not subject to copyrights in USA.

1. edition, March 2016

Copyright: National Food Institute, Technical University of Denmark

Photo: Colourbox

ISBN: 978-87-93109-78-0

The report is available at http://www.globalmicrobialidentifier.org/Workgroups

National Food Institute

Technical University of Denmark

Søltofts Plads

Building 221

DK-2800 Kgs. Lyngby

Denmark

Tel: +45 35 88 70 00

Fax +45 35 88 70 01

Contents

List of Abbreviations 5
1. Introduction 6
2. Materials and Methods 7
2.1 Participants 7
2.2 Strains 7
2.3 Genomes 8
2.4 Distribution 8
2.5 Procedure 8
3. Results 11
3.1 Participation 11
3.2 Method description of the Wet lab component 12
3.3 Sequencing, Wet lab – MLST, and antimicrobial resistance genes 13
3.3 Sequencing, Wet lab – Quality markers 15
3.3 Sequencing, Wet lab – Phylogeny 16
3.4 SurveyMonkey, Dry lab 17
3.5 Sequence Matrix and Phylogeny, Dry lab 17
4. Discussion 18
4.1 Overall 18
4.2 SurveyMonkey and Sequencing, Wet lab 19
4.3 SurveyMonkey and Phylogeny, Dry lab 20
5. Conclusions 21
6. Acknowledgement 21
Appendices 22
Reference List 23

List of Abbreviations

AMR	**Antimicrobial resistance**
CC	**Clonal complex**
CGE	**Center for Genomic Epidemiology**
DDBJ	**DNA Data Bank of Japan**
DNA	**Deoxyribonucleic acid**
DTU	**Technical University of Denmark**
EBI	**European Bioinformatics Institute**
E. coli	***Escherichia coli***
GMI	**Global Microbial Identifier**
IATA	**International Air Transportation Association**
NCBI	**National Center for Biotechnology Information**
MLST	**Multi Locus Sequence Typing**
PT	**Proficiency test**
QC	**Quality Control**
SNP	**Single Nucleotide Polymorphism**
S. aureus	***Staphylococcus aureus***
TAG	**Technical Advisory Group**
USA	**United States of America**
US FDA	**U.S. Food and Drug Administration**
WG	**Working group**
WGS	**Whole genome sequencing**

1. Introduction

The main objective of this pilot proficiency test (PT) is to facilitate the production of reliable laboratory results of consistently good quality within the area of whole genome sequencing (WGS).

Initially, a survey was launched to ensure that a future PT would serve the prober target audience as well as the bacterial pathogens of interest. In addition, the survey captured the information about what current quality markers being employed to ensure high quality sequencing data (7). The results of this survey were utilized to create the foundation of this pilot PT.

Specifically, the PT will evaluate the consistency and robustness of Global Microbial Identifier (GMI) member's and other's ability to perform deoxyribonucleic acid (DNA) extraction, library preparation, the WGS, the assembly and phylogenetic analysis following different laboratory protocols, software tools, and sequence platforms for the reliability of submitted sequence data. This ensures harmonization and standardization in WGS and data analysis, with the aim to produce comparable data for the GMI initiative. A further objective is to assess and improve the process of uploading WGS data to databases such as National Center for Biotechnology Information (NCBI), European Bioinformatics Institute (EBI) and DNA Data Bank of Japan (DDBJ). To meet these objectives, the laboratory work and analyses performed for this PT should be performed using the methods routinely employed in the individual laboratories.

The PT consists of two "Wet-lab" and one "Dry-lab" component targeting various bacterial pathogens. The Wet-lab components assess the laboratories ability to perform DNA preparation, sequencing procedures and, if laboratories routinely do so, the analysis of epidemiological markers; Multi Locus Sequence Typing (MLST) and antimicrobial resistance (AMR) genes. The Dry-lab component captures the laboratories analysis of a WGS dataset to distinguish between clonally related genomes.

The results presented in this report were part of a pilot PT that was initiated as a 'test run' to allow for collection of experience with the purpose of adjusting documentation and/or the practical approach before the full roll-out of the PT, which is scheduled to be launched to a global audience in 2015.

The main organizers of the GMI pilot PT are Technical University of Denmark (DTU), Kgs. Lyngby, Denmark in collaboration with U.S. Food and Drug Administration (US FDA), Silver Spring, Maryland, United States of America (USA). The Technical Advisory Group (TAG) for the GMI PT program consists of members and institutions of working group (WG) 4. The GMI PT organizers strive towards conducting the PT annually.

Individual laboratory data are confidential and only known by the participating laboratory, the PT organizers (DTU Food, US FDA), and potential assisting members of the TAG. All summary conclusions are made public. The tentative goal set by the GMI PT organizers and TAG aim towards having all participating laboratories performing WGS on single bacterial isolate cultures and supplied bacterial DNA allowing the TAG to set future thresholds for Quality Control (QC) as well as identifying all related genomes in the cluster analysis based on supplied data files.

2. Materials and Methods

2.1 Participants

A pre-notification (App. 1) to announce the GMI pilot proficiency test was distributed on the 17th June 2014 by e-mail to the 11 invited participants. The TAG agreed that participants of the pilot PT should all be current WG4 members to allow for collection of experience with the purpose of adjusting documentation or the practical approach before the full roll-out of the PT. Institutes signing up and submitting results were from Australia (1), Denmark (2), Germany (1), Spain (1), and USA (3). One country had intended to participate but reconsidered due to the political situation in the country at the time. Of the eight participants, seven delivered results for the Wet lab component, and another seven delivered results for the Dry lab component. For both components, some of the participating laboratories delivered only parts of the requested results and/or information.

The results from the participating laboratories are presented and evaluated in this report; i.e. results from eight laboratories representing five countries and three continents are included.

In addition to reporting results as requested in the PT protocol (Appendix 2), the participants were asked to capture and report any unfinished details that might still be in the PT material, e.g. in descriptions or in the handling/organization. This was conducted with the purpose of making the relevant adjustments before launching the full roll-out of the PT.

2.2 Strains

Two strains of *Salmonella* spp, *Escherichia coli (E. coli)*, and *Staphylococcus aureus* (*S. aureus*) were selected for the Wet-lab of the GMI PT in 2014. The GMI14-001 was the *S.* Typhimurium strain LT2, the principal strain for cellular and molecular biology in *Salmonella* that was isolated in the 1940s (8). The second *Salmonella* strain; GMI14-002, was a *S.* Concord harbouring the antimicrobial resistance gene; $bla_{\text{SHV-12}}$ gene and originally recovered from Ethiopian adoptees (3). The rationale for inclusion of the latter strain was our finding of the lack of amplification of the resistance gene on one specific sequencing platform. An O139 *E. coli* strain (GMI14-003) associated with edema disease in pigs (1) was intended to be included the panel. This strain and the corresponding purified DNA was, however, due to contamination issues, excluded before the shipment of the PT-material to the participants. In addition, an *E. coli* culture (GMI14-004) of unknown MLST obtained from a Danish pig isolated in 2011 was also included. The two *S. aureus* isolates included belonged to the clonal complexes (CC); CC30 (GMI14-005) and CC9 (GMI14-006) and isolated in 1955 and 2011 from a pig and cattle in Denmark, respectively.

Individual sets of the strains were lyophilized as KWIK STIKs by Microbiologics, St. Cloud, Minnesota, USA and the corresponding DNA were purified and pooled by DTU-Food prior to distribution in individual vials for each participant.

To better be able to assess the differences in the sequences generated by the participants, each of the six strains in the Wet-lab component were sequenced on the PacBio to get a closed reference genome. This was done by creating 10KB template libraries using “10kb DNA Template Prep Kit 1.0” from Pacific Biosciences, which were then sequenced using C2 chemistry on single-molecule

real-time (SMRT) cells with a 180min collection protocol. The data was then de novo assembled using the Hierarchical Genome Assembly Process (HGAP) within the Pacific Biosciences SMRTAnalysis software package. Polishing and finishing the genome were performed with custom python scripts, Quiver and Gepard, a dot plot tool to identify overlapping regions.

2.3 Genomes

A dataset consisting of raw sequence files (fastq files produced with Illumina MiSeq benchtop sequencer) were constructed for each of the three taxonomic groups represented in the Wet lab portion of the pilot-PT (*E. coli*, N = 22; *S. aureus*, N = 24; and *S.* Typhimurium, N = 20). Throughout this document this component of the pilot PT is referred to as the Dry lab. The samples within each dataset were chosen to generally represent the degree of variation that might be found in a typical outbreak/trace-back investigation within which there are a number of samples that differ by only a few Single Nucleotide Polymorphism's (SNPs) with the remainder of the samples being quite genetically distinct (e.g., > 100 SNP differences).

2.4 Distribution

On 26 June 2014, for the Wet lab component of the PT, bacterial strains in agar stab cultures together with the corresponding purified and dried DNA and a welcome letter (App. 3) were dispatched in double pack containers (class UN 6.2) to the participating laboratories according to the International Air Transport Association (IATA) regulations as UN3373, biological substances Category B. On the same date, participants in the Dry lab component received the protocol in an email. An appendix to the protocol informed of how to access the sequence data to be analyzed. This data could be downloaded from a password protected ftp-server which was accessible for all participants.

2.5 Procedure

The protocols and all relevant information was sent by email and subsequently uploaded on the GMI website for direct download (http://www.globalmicrobialidentifier.org/Workgroups/About-the-GMI-Proficiency-Test-Pilot-2014), thereby PT participants could access necessary information at any time.

The protocol presented instructions as to the handling of the received bacterial cultures and DNA and participants were instructed on how to obtain access to the fastq datasets to be included in the Dry lab component.

Participants were requested to capture information in relation to the questions presented in the SurveyMonkey, i.e. one for bacterial cultures and DNA, and one for the fastq datasets, for the components relevant to each participant's level of participation.

With the aim to improve future PT trials, participants were also encouraged to communicate directly with the PT-coordinator or to collect any questions or comments to the set-up or the documentation provided at the pilot PT.

Deadline for submission of results was set for 15th August 2014, and after this date, participants who had not yet submitted results according to the level of their sign-up, were approached to confirm if they were planning on submitting results. By the end of August 2014, all relevant data was captured and the data analysis was instigated. This report summarizes the results and allows for ensures full anonymity for the participants, as only the PT-organizers has access to the individual results.

2.5.1 SurveyMonkey, Wet lab

Apart from three questions relating to the contact information of the participant, thirty-two questions were asked of those participating in the Wet lab component. The questions focused on the storage of bacterial cultures and DNA prior to analysis, the cultivation and DNA extraction procedure, the quality assurance parameters applied, details related to the sequencing and analysis of the obtained sequencing data (App. 2 (appendix 2 in the PT-protocol)).

2.5.2 Sequencing, Wet lab

The participants in the Wet lab analysis uploaded raw sequence files in fastq format. The reads were de novo assembled applying the standard assembly pipeline used by the web-services from Center for Genomic Epidemiology (CGE) https://cge.cbs.dtu.dk//services/all.php, except for the reads which were not trimmed prior to the assembly.
For the raw reads, the following QC metrics were calculated:

- Total numbers of reads (for paired-end reads, the total numbers of reads is calculated as the sum of reads in the two files)
- Average read length (bp)
- Number of reads that map to the total reference DNA (chromosome + any plasmids) using BWA (6)
- Proportion of reads that map to total reference DNA
- Number of reads that map to reference chromosome
- Proportion of reads that map to reference chromosome out of all reads that map to total reference DNA
- Number of reads that map to reference plasmid #1
- Proportion of reads that map to reference plasmid #1 out of all reads that map to total reference DNA
- Coverage, total reference DNA. The number of reads mapping to the total reference DNA multiplied with the average length of the reads divided by the total size of the reference genome

- Coverage, reference chromosome
- Coverage, reference plasmid #1

For the assemblies, the following QC parameters were calculated:

- Total size of assembly (bp) (all contigs)
- Total number of contigs
- Number of contigs with a length above 200 bp
- N50 (defined as the length of the shortest contig, in the set of largest contigs that represents at least 50% of the assembly)
- NG50 (defined as the length of the shortest contig, in the set of largest contigs that represents at least 50% of the reference genome)
- Coverage (The sum (lengths) of "islands") for total DNA calculated using COMPASS (2)
- Validity for total DNA which reflects the alignable or validatable fraction of the assembled sequence (2)
- Multiplicity for total DNA which reflects the ratio of the length of alignable assembled sequence to covered sequence of the reference (2)
- Parsimony for total DNA which reflects how many bases of assembled sequence need to be inspected to find one base of real, validatable sequence (2)

In addition to the calculation of the above QC metrics and parameters, participants were requested to provide the identification of the strains corresponding MLST and AMR genes to support the assessment of the sequence quality. Participates identified the MLSTs and AMR genes using the software of their choice. To assess the proficiency of the participants, the PT organizers used a command line version of the CGE MLST-Finder v.1.7 (5) and ResFinder 2.1 (9) (Threshold for %ID = 98% and HSP/Query length = 60%) including the CGE standard assembly pipeline on the participant's raw reads to compare the results with those reported by the participants. Furthermore, strain specific reference routed phylogenetic SNP trees were created (4) using the raw reads of both the culture and corresponding DNA submitted by each of the participants. This will support the assessment of the sequence quality of the participants.

The phylogenetic SNP trees were created using the pipeline; CSI phylogeny v.1.0a available from CGE. The paired-end reads were mapped to the reference genome using Burrows-Wheeler Aligner (BWA) version 0.7.2. The depth at each mapped position was calculated using genomeCoverageBed, which is part of BEDTools version 2.16.2. SNPs were called using 'mpileup'

module in SAMTools version 0.1.18. SNPs were filtered out if the depth at the SNP position was not at least 10x or at least 10% of the average depth for the particular genome mapping. Subsequently, SNPs were selected when meeting the following criteria: 1) a minimum distance of 15 bps between each SNP, 2) the mapping quality was more above 20, 4) the SNP quality was more than 20 and 5) all indels were excluded.

The qualified SNPs from each genome were concatenated to a single alignment corresponding to position of the reference genome. The concatenated sequences were subjected to parsimony tree construction using PhyML with HKY85 substitution model and 100 bootstrap replicates.

2.5.3 SurveyMonkey, Dry lab

Twenty-one questions were asked of those participating in the Dry lab component, which focused on the preprocessing of fastq files that may occur within a lab in addition to the methods that were used to construct a cluster for outbreak and trace-back investigations (App. 2 (appendix 3 in the PT-protocol)).

2.5.4. Sequence Matrix and Phylogeny, Dry lab

The participants in the Dry lab component were asked to provide the results that they would typically produce as part of an investigation/study. In order to increase the number of laboratories willing to participate and get a more accurate picture of the diversity of methods being employed, we did not specify that any particular analysis pipeline or software had to be used. However, to provide some guidance as to what we were looking for we asked participants to upload, if they created them, a DNA/SNP matrix, a phylogeny, contigs, and distance matrix. The 21 questions that were part of the SurveyMonkey for the Dry lab compliment the files that were uploaded and provide additional information about the analyses performed by each participant (e.g., reference or reference-free SNP detection, MLST, etc. (App. 2 (appendix 3 in the PT-protocol)).

For each SNP matrix we plotted the average pairwise distances among all samples against the size of the matrix to illustrate how the pipelines used by the different labs differed in the number of SNPs extracted (length of matrix) and the information content (average pairwise difference).

Analyses of the topologies were complicated by the fact that participants differed in the number of samples included, which makes it difficult to conduct topological comparisons. Rather than drop tips to make the datasets be consistent in the samples they contained, we have not analyzed the topologies but suggest that for the future roll out the samples included in the matrix be only those samples provided (i.e., if a reference was used that was not part of the dataset, do not include that in the final results submitted).

3. Results

3.1 Participation

A total of nine laboratories responded to the pre-notification and were enrolled in the GMI pilot PT. When the deadline for submitting results was reached, eight laboratories in five countries had

uploaded data. The following countries provided data for at least one of the PT components (Figure 1): Australia, Denmark (2), Germany, Spain, and USA (3). Seven of the participating laboratories were from the public health sector and one laboratory represented food safety sector.

3.2 Method description of the Wet lab component

The laboratories received the reference material in the period from 24th of June to the 8th of July 2014 and uploaded the results from the 29th of July to the 25th of August 2014.

The bacterial cultures were stored at 4°C by 50% (n = 3) of the participants prior to the analysis. In addition, two participants (33%) stored the reference material at room temperature and one participant started the analysis on arrival of the organisms.

Three participants (50%) stored the DNA in the time between reception and processing at room temperature whereas the remaining part of the participants stored the DNA at either -20°C, 4°C or started the analysis instantly.

All six participants inoculated the bacterial cultures onto various types of blood agar and incubated the plates at 37°C between 16 to 24 hours.

The Genomic DNA was extracted from both the Gram negative and positive strains using the Qiagen DNeasy Blood and Tissue kit according to the Gram negative and positive protocol by four participants (67%) whereas the two remaining participants used either the Invitrogen EasyDNA kit or the QIAamp DNA MiniKit. Five of the participants have modified the used Gram negative protocols by lysostaphin treatment prior to extraction and reduced enzyme incubation time from 3h to 1h. In contrast, four of the participants have modified the Gram positive protocol including a lysostaphin treatment step.

DNA concentrations (ng/µl) of the bacterial cultures were determined prior to library preparation on a Qubit by four participants (67%) whereas the remaining two participants used either Nanodrop or Picogreen (Figure 2A). For the DNA received, five (83%) participants used the Qubit to measure the DNA concentration (ng/µl) prior to library preparation. One used the Picogreen (Figure 2B).

Five (83%) participants measured the DNA concentration including the total amount of the five bacterial cultures whereas all six participants measured the DNA concentration of the provided DNA reference material (Table 1 and 2).

For the *Salmonella* cultures, the DNA concentration ranged from 14.1 to 65.2 ng/µl and from 0.2 to 128 ng/µl for the provided DNA (Table 1). For the *E. coli* culture the DNA concentration ranged from 9.64 to 110 ng/µl and from 0.2 to 80.7 ng/µl for the DNA. The DNA concentration ranged from 9.48 to 216 ng/µl and from 0.2 to 147 ng/µl for the *S. aureus* bacterial culture and DNA, respectively. One laboratory reported a concentration of 0.2 ng/µl for all DNA samples (Table 1).

For the *Salmonella* culture, the total amount of DNA ranged from 1.2 to 6.5 µg and from 0.02 to 6.4 µg for the provided DNA (Table 2). For the *E. coli* culture the total amount of DNA ranged from 1.9 to 11 µg and from 0.02 to 7.2 µg for the total amount of DNA. The total amount of DNA ranged from 0.7 to 43.2 µg and from 0.02 to 7.9 µg for the *S. aureus* culture and DNA, respectively. One laboratory reported a concentration of 0.02 µg for all DNA samples (Table 2).

Six and five participants responded to the method applied to measure the DNA quality (e.g. RIN, 260/280 ratio and/or 260/230 ratio) prior to library preparation for bacterial cultures and DNA received. For bacterial cultures, three (50%) of the laboratories used the Nanodrop whereas one of the remaining three participants assessed the quality visually on a agarose gel. Two (33%) did not measure the DNA quality (Figure 3A). For the DNA received, one (20%) participant used the Nanodrop for measuring the DNA quality prior to library preparation and two (40%) either the Picogreen or visual assessment. One participant did not measure the DNA quality of the received DNA prior to library preparation (Figure 3B).

Three and two participants reported the measurement of the DNA quality (e.g. RIN or 260/280 ratio) for bacterial cultures and DNA received (Table 3). Among the three laboratories providing data of the DNA quality for the cultures, the level ranged from about 1.7 to 2.08. For the received DNA, the level between the three organisms was consistent around 2.0 (Table 3).

Two participants reported the measurement of the DNA quality (260/230 ratio) for bacterial cultures and DNA received (Table 4). For the cultures and received DNA, the DNA quality was consistent around 0.5 to 2.0 (Table 4).

Of the six participants, four used the Illumina Nextera XT DNA sample preparation kit FC-131-1096 and one indicated using the Illumina NEB Next Ultra DNA Library prep kit for the preparation of the sample library before sequencing. One of the four participants using the Illumina Nextera XT DNA kit indicated using this in combination with the Nextera XT Index kit FC-131-1002. In addition, one participant used the Ion Xpress Plus Fragment Library kit 4471248.

The genomic DNA was prepared for pair-end sequencing by five (83%) participants whereas one prepared for single-end sequencing. The libraries were sequenced by five participants (83%) using an Illumina MiSeq platform and one used the Ion Torrent PGM platform (Figure 4). The read length of the sequences was set between 150 (n = 1), 200 (n = 1), 250 (n = 1) and up to 300 bp (n = 3). The reads were by five of the six participants trimmed before upload. Four participants indicated that if assembled by themselves, two would have used Velvet (https://www.ebi.ac.uk/~zerbino/velvet/), one would have used the Assemble pipeline (version 1.0) available from CGE and the last participant would have used the CLC Genomics Workbench 7.

3.3 Sequencing, Wet lab – MLST, and antimicrobial resistance genes

Five participants determined the MLST-type or alleles and detected AMR genes from both the received bacterial culture and corresponding DNA The four Miseq users also used the CGE "ResFinder" tool either solely or in combination with other online databases or in house scripts to detect the AMR genes. The Ion Torrent user utilized an in house tool; custom pseudomolecules (Table 5).

Unfortunately, the WGS data revealed that the culture and the corresponding DNA of GMI14-004-*E.coli* were contaminated. Up to approximately 15% of the sequence reads from this strain did not map to the reference DNA regardless of which participant has submitted the data why the strain has been analyzed with care.

For *Salmonella*; GMI14-001 the expected MLST was ST19 which were found in most cases. The expected results for participant 2 deviated for both the culture and the DNA with six alleles using the CGE reference MLST tool. The same was observed for participant 4 testing the culture using the CGE reference tool resulting in one deviating allele. In contrast, participant 9 seemed to have reported for the culture and the DNA sample using own tool the correct allele numbers but displaced in relation to the locos. The same displacement was also observed for the other PT strains by participant 9.

The resistance gene; *aac*(6')-*1aa* was reported by all participants for both the culture and DNA using own tools except for participant 3 who did not report any resistance genes. In contract, using the CGE reference tool; ResFinder, all participants found the related gene; *aac*(6')-*Iy* (Table 6).

The MLST 599 was expected in *Salmonella* strain; GMI14-002. This was reported by all participants except for participant 9 who for both the culture and DNA using an in-house tool reported a similar phenomenon as for GMI14-001. It is clear that this is a mistake and not related to the used tool, sequencing or reference materials as the allele number for *aro*C correspond to the MLST.

Some variation among the participants, the reference material and tools used were observed for the detection of AMR genes in in *Salmonella* strain; GMI14-002. Concordance between reported genes detected by own tools and the CGE reference tool and between culture and DNA samples was observed for participant 1 except for reporting the presence of *cat*A2 using own tool. In addition, using own tool on the DNA, an absence of *aac*(6')-*IIc* was observed.

Less concordance between reported AMR genes detected by own tools and the CGE reference tool was observed for participant 2. Participant 2 did not report using own tool in both samples the presence of *dfr*A18, *ere*(A), *qnr*B49, *str*B, *tet*(D), and *aac*(6')-*Iy* in comparison to others. However, these AMR genes were observed using the CGE reference tool.

Concordance was observed between culture and DNA and between tools for participant 4 except for the reported presence of *cat*A2 and *aac*(6')-*IIb* using own tool. These genes were infrequently reported by other participants using own tools.

For participant 9, concordance was observed in most cases except for reporting the presence of *cat*A2 and *Qnr*B2 and absence of *flo*R, *Qnr*B49, and *tet*(A) by own tool compared to the CGE reference tool (Table 8).

Unfortunately, no MLST has been assigned to *E. coli* strain; GMI14-004. However, the excepted allele code was 233, 2, 29, 167, 4, 16, and 4. This appeared to create some confusion as the majority of the participants reported this MLST incorrectly or did not report the alleles. It was evident that participant 9 had a problem reporting the MLST as the phenomenon mentioned above was consistent for also the *E. coli* samples. Participant 4 reported for both the culture and DNA samples using own tool all seven alleles identical but incorrect. This counted also for participant 3 but only for loci *rec*A.

The *E. coli* strain; GMI14-004 was a pan-susceptible strain harboring no antimicrobial resistance genes. Despite of this, the presence of *qnr*B49, *str*A, *sul*1, *sul*2, and *aac*(6')-*IIc* were observed using the CGE reference tool for participant 2 in the culture. Additional AMR genes were both reported

and observed using own and the CGE reference tool on the DNA sample. This supports the above mentioned contamination as also the alleles for MLST were to some degree incorrect (Table 10).

For *S. aureus*; GMI14-005 the expected MLST was ST433 which were found in most cases except for participant 2 and 9. Participant 9 also for *S. aureus* has the above mentioned problem reporting the MLST. For participant 2 all alleles were observed to be incorrect using the CGE reference tool on the culture but correct using own tool.

The reported and observed AMR genes using own and CGE reference tool were identical for both the culture and DNA samples. The *str*2 was not reported by participant 1 but observed using the CGE reference tool. Interestingly, participant 2 reported the presence of *bla*Z, *mec*A(10), *str*2, and *tet*(38) using own tool in the culture and the DNA except for *str*2. In contrast, only the *mec*A(10) gene was observed using the reference CGE tool. Participant 3 reported the presence of the *vga*(A) and *czr*C genes for both the culture and the DNA samples which was not observed using the CGE reference tool. In contrast, the *tet*(38) and *str*2 were detected using the CGE reference tool. Using both tools also allowed for the detection of the *mec*A and *bla*Z genes. The AMR profile for participant 4 was identical for both the culture and DNA samples using both tools with the exception of additional observation of the *tet*(38) and *str*2. Similarly, consistency was observed for participant 9 where the additional observations were related to the reporting of *fus*A5 and *aac*(3)-Ik genes by own tool (Table 12).

The MLST 9 was expected in *S. aureus* strain; GMI14-006. This was reported and observed by most of the participants for both sample types using both own and the CGE reference tool. However, participant 9 encountered the same problem also for this stain assigning the MLST as described previously using own tool. No problems were observed using the reference tool. In contrast, participant 2 reported the correct MLST for both the culture and DNA samples using own tool but failed when using the CGE reference tool.

Participant 1, 4 and 9 all reported using own tool and observed when using the CGE reference tool the presence of *tet*(38) in both sample types. In addition, participant 4 reported using own tool the presence of the *nor*A gene and participant 9 reported the *fus*A6 and *bla*Z genes. Participant 3 reported using own tool for both sample types the presence of the *fos*B gene whereas the *tet*(38) gene was observed when using the CGE reference tool. The AMR profile for participant 2 seemed different compared to the other data reported. Participant 2 reported using own tool the presence of *bla*Z, *mec*A, and *tet*(38) genes in both the culture and DNA samples. This profile was similar to the expected profile of *S. aureus* strain; GMI14-005. Using the CGE reference tool on the culture provided the same profile as reported using own tool. However, using the reference tool on the DNA sample revealed the presence of *str*A and *str*B genes (Table 14).

3.3 Sequencing, Wet lab – Quality markers

Six participants submitted data related to the quality metrics and parameters from both the received bacterial culture and corresponding DNA.

Initially, the quality markers were evaluated for potential contamination or a low performance by accessing the “total size of assembly”, “percentage of total size of assembly per total size of the reference DNA”, “N50” and “depth of coverage”. The quality markers for all PT strains produced

by participant 2 were all outliers with the exception of GMI14-002-Salmonella. The values for "percentage of total size of assembly per total size of the reference DNA" were all considerable above 100% indicating the submitted data being contaminated (Table 7, 11, 13, 15). In contrast, participant 5 obtained only 24.4% of total size of assembly per total size of the reference DNA indicating a poor sequencing of the DNA sample 4.

The percentages of total size of assembly per total size of the reference DNA ranged from 93.6% (participant 5, culture GMI14-005-S.aureus (Table 13)) to 104.1% (participant 9, DNA GMI14-004-E.coli (Table 11)) with strain medians of 99.1%, 98.3%, 102.4% (GMI14-004-E.coli), 97.9%, and 98.0% (Table 7, 11, 13, 15). As mentioned above, the samples of GMI14-004-E.coli were 15% contaminated but was included as an example of low contamination.

The number of reads seemed to be consistent across the strains per participant (excluding poor and contaminated data from participant 2) ranging from 883248 (participant 3, GMI14-001-Salmonella) to 7243770 (participant 4, GMI14-005-S.aureus) with strain medians of 2334844, 1900294, 2053210 (GMI14-004-E.coli), 2380956, 2532644 (strain 1-6) (Figure 6).

The depth of coverage (x) also seemed to be consistent across the strains per participant (excluding poor and contaminated data from participant 2) with lower coverage of (GMI14-004-E.coli) (median of 59x) compared to the *S. aureus* strains which in general provided a higher coverage; median of 151x and 140x (Figure 7).

The total number of contigs seemed likewise to be quite consistent per strains and participants but with more variation across the PT strains with a higher number of contigs for (GMI14-004-E.coli) (median of 363) compared to the *S. aureus* strain GMI14-006-S.aureus which in general provided a lower number of contigs; median of 104 (Figure 8).

The submitted data of N50 were highly similar per strain across participants i.e. from 40121 to 40943 for strain GMI14-001-Salmonella (Table 7). However, more variation was observed across the "species" with medians for *E.coli* and *Salmonella* of 40872, 43232, and 44730, respectively compared to medians of *S. aureus*; 23816 and 22217, respectively (Figure 9, Table 7, 11, 13, 15).

3.3 Sequencing, Wet lab – Phylogeny

It was not possible to create phylogenetic trees for strain GMI14-001-Salmonella, GMI14-002-Salmonella and GMI14-004-E.coli as no SNP differences between the participants' submitted genome data were observed.

In contrast, a total of 18244 SNPs difference was observed between culture and DNA samples of GMI14-005-S.aureus from participant 1. As no SNPs were observed among the other GMI14-005-S.aureus genomes submitted by the other participants, suggesting that culture of GMI14-005-S.aureus from participant 1 was contaminated (Figure 5).

A minor SNP variation of eight SNPs was observed in the culture of GMI14-006-S.aureus submitted by participant 2 compared to the other genomes of strain 6 (Figure 6). This can be explained by the contamination described above.

3.4 SurveyMonkey, Dry lab

Eight laboratories participated in the Dry lab component of the pilot PT within which three datasets consisting of raw fastq files were provided. Participants were asked to carry out the analyses they would typically perform as part of an outbreak investigation or study.

Quality filtering, assembly and reference genomes:

Half the of the participants reported that they did quality filter raw reads; two laboratories reported using Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic), one used cutadapt (http://code.google.com/p/cutadapt/), and the fourth used NGS QC Toolkit (http://www.nipgr.res.in/ngsqctoolkit.html).

Two of the participants also stated that they filtered contigs below a certain length (<200 and <1000); the other six labs did not report doing any filtering of contigs based on length.

Six of the labs replied that they did calculate N50 two of which did so after filtering. Six labs also reported estimating coverage and the methods for doing so included using the expected genome size, as part of the pipeline used to quality filter and assembly (i.e., NGOPT), and as part of the mapping process.

Two of the eight labs did not perform assemblies and for those that did, Velvet was the dominant assembler used (Table 16).

All eight labs reported using a reference genome within their pipeline and interestingly not all labs chose the same reference (Table 17).

Species verification and typing:

Four labs reported not attempting to verify the species but those that did used Kraken (http://ccb.jhu.edu/software/kraken/), pathoscope (http://sourceforge.net/projects/pathoscope/), and Kmer Finder (https://cge.cbs.dtu.dk/services/KmerFinder/). Interestingly, labs reported being able to identify the species to which the fastq files belonged for sometimes only a specific group, which was likely an artifact of the setup in SurveyMonkey (i.e., only one answer was allowed and thus why one one participant replied “yes for all S. Typhimurium” but not the other groups). The SurveyMonkey has been edited to address this issue.. Also of note is that one sample of E. coli was found to be contaminated with *Morganella morganii*, which is correct. This sample has been removed from the dataset that will be provided during the full roll out.

Four participants responded that they performed MLST.

3.5 Sequence Matrix and Phylogeny, Dry lab

All participants in the Dry lab stated that they used a SNP matrix for inferring phylogenies rather than, for example, a gene-based approach such as whole genome MLST. However, the survey was not adequately designed to extract the details of how the SNP matrix was constructed, which will be addressed in the full roll out of the PT.

Perhaps one of the more interesting results from the Dry lab component was that the number of sequences in the matrix/tips in the tree differed among the participants as some included a reference that was not part of the original dataset. This makes it extremely difficult to compare matrices and topologies and thus a subset of the potential comparative analyses were performed for the pilot.

The matrices all differed in length with a median of 13196 bp (Table 18; Fig. 10). However, within *E. coli* the range was exceptionally large with one participant submitting a matrix that was approximately three times larger than the next biggest matrix.

Although the participants produced different sized matrices, there is some consistency in the relative differences among samples. For example, within *E.* coli there are a number of samples that are consistently more different from other samples than the majority of samples (Fig. 11). This consistency in relative differences also suggests that the different methods may provide similar topologies and, thus, there may be congruence among conclusions based on the topology. However, we have not tested the differences in topologies due to different numbers of samples within the trees.

All participants also stated that they inferred a phylogeny using a maximum-likelihood approach where the actual program used included PhyML (N = 2), RAxML (N = 3), GARLI (N =1), FastTree (N = 1), and Geneious (N = 1). Geneious software package includes a number of likelihood approaches so it is difficult to determine exactly which one was used.

The participants differed as to how many clusters (2 to 6) were on the tree and the level of detail describing the process used to identify them (Table 19; Appendix 4).

4. Discussion

4.1 Overall

The pilot PT was useful as it indicated the critical points where an improvement will be necessary for the full PT roll out. It provided some confusion and extra work to sort the submitted genomes by strains, source and participants mainly due to participants naming the sequence data in various ways. In the full PT roll out, the PT organizers will modify the protocol to ensure a proper way naming the sequence data to better distinguish data between the participants.

In addition, the PT organizers observed some problems related to the submission to the ftp site which we will re-solve by submitting the Dry lab part to a ftp site but the Wet lab part via a CGE batch upload tool which will create an instant feedback of the quality marker results.

Unfortunately, PT organizers was not aware that one of the *E. coli* strains included the PT was not designated a MLST at the time of this exercise. PT organizers will ensure to have all data collected prior to the full PT roll out. In addition, the fact that the PT strain; GMI14-004-E.coli was 15% contaminated prior to dispatching the samples to the participants was discovered too late by PT organizers.

It was not possible to conduct the analysis as expected for the Dry-lab part mostly related to missing information. To adjust this, the PT organizers will modify the protocol, the survey to capture vital non-sequencing data and the method for analysis for the full PT roll out.

4.2 SurveyMonkey and Sequencing, Wet lab

The majority of the submitted MLST data were correct and in line with the expected value. The results of MLST analysis revealed a systematic error for participant 9 either when submitting the data. The reported allele numbers were for all loci dislocated by one allele "downward" with the correct MLST reported for allele *aro*C using own tool. However, the MLST was correct for all PT strains when re-analysed using the reference method. The MLST results also revealed some deviations for participant 2 which compared to AMR genes and the quality markers showed that the whole genome sequencing at some point of stage for all PT strains except for GMI14-002-Salmonella had been contaminated. A few other deviations were observed i.e. *his*D in the culture of GMI14-001-Salmonella which deviated for participant 4 between using own tool and the CGE reference assembler. Interestingly, participant 4 also used the CGE reference assembler and MLST tool as "own tool" why it is suspected that the deviations might be a results of running the tools as the online stand-alone version compared to running the analysis by command line. Another example of deviations was observed for participant 3 testing the PT strain GMI14-004-E.coli for both sample types. Running the analysis using own tool resulted in an incorrect allele number; 99 for *rec*A whereas a different incorrect allele number; 215 for *rec*A was reported using the CGE reference tool. This might be a result of the assembler used which potentially could have resulted in truncated contigs.

Most of the submitted AMR genes were in concordance with the expected results. However, some deviations were observed. All participants; expect for participant 3, despite of own tool used reported the AMR gene; *aac*(6')-*Iaa* for both sample types whereas the gene reported using the CGE reference tool was *aac(6')-Iy.* The reason for the discrepancy was related to a change in the CGE reference tool database; "the ResFinder" where the two genes were re-named. This affected all participants using the ResFinder tool.

Participant 3 reported using an in-house AMR gene tool; "pseudomolecules" which potential could be the reason for different AMR profiles reported compared to the other profiles in concordance with the expected i.e. GMI14-002-Salmonella. Similarly, participant 2 also reported different AMR profiles which were not surprising taking the contamination into account. It could be speculated if the contamination in PT strain GMI14-004-E.coli could be a result of mixing with also GMI14-002-Salmonella. The PT strain GMI14-004-E.coli was expected being pan-susceptible but reported by participant 2 harbouring a highly similar AMR profile as the PT strain; GMI14-002-Salmonella.

The phylogenetic analysis revealed no SNP difference between the participants' genomes for PT strains; GMI14-001-Salmonella, GMI14-002-Salmonella, and GMI14-004-E.coli. Only a few; eight SNPs were observed for the culture of GMI14-006-S.aureus submitted by participant 2. Eight SNPs from clonal strains is a bit more than expected in relation to spontaneous mutations why the discrepancy more likely was related to the mentioned contaminations. In contrast, a huge difference in number of SNPs was observed creating the phylogenetic tree of GMI14-005-S.aureus with 18,244 SNPs between the culture and DNA submitted by participant 1. It was not possible to detect any other deviations e.g. for MLST, AMR genes and quality markers.

One of the objectives for the GMI PT was to assess a range of quality markers enabling the organizers to propose / set quality control thresholds. However, this was not possible taken into

account the few participants in this pilot PT. Nonetheless, it was possible to detect contamination of the PT strains from participant 2 based on primarily; “Total size of assembly”, “Percentage of total size of assembly per total size of DNA”, “Total no. of contigs”, “No. of contigs > 200 bp”, and “N50”. N50 is one of the vital quality markers and good quality has been suggested to be around 30,000 across all bacterial species. It was surprising to see that N50 seemed to be that species depended as all N50’s from all participants clustered per PT strains and species. By expanding the PT to a greater number of participants with a full roll out of a future PT, the PT organizers hope to be able to set some quality control thresholds ensuring reliable sequencing data for the future.

4.3 SurveyMonkey and Phylogeny, Dry lab

The survey of the Dry lab component highlighted the diversity of methods that exist to analyse and cluster based on whole genome sequence data. For example, not all labs quality filtered reads before performing down stream analyses and not all labs tested for contamination or verified species. There was some consistency in that each lab used a reference-based approach to identify variant sites but at the same time each lab often used a different reference.

The participants also differed in what samples were included in the final matrix, which made it difficult to compare topologies.

The results from the pilot PT also highlighted some deficiencies in the survey that inhibited our ability to adequately compare the methods of the different labs. In particular, no questions were posed regarding the software used to perform mapping and identify variant sites. Also, the question regarding how clusters were determined was vague and as a result the answers were difficult to interpret; perhaps that question should be removed.

Below are some specific recommendations/comments for the full PT based on the results from the pilot:

- Clearly state how sequences are to be named in the fasta file and subsequent tree
- Instruct participants to only include samples that were supplied (e.g., if they chose a reference outside the provided samples do not include that reference in the SNP matrix or tree file)
- If a reference-based approach was used ask the following:
    - What mapper was used
    - What software was used to identify variant sites
- If a reference-based approach is used then they can ignore the questions about assemblies. Perhaps the survey could be setup to first ask if they use an assembly or reference based method and depending on the answer to that question the participant moves to a certain section were relevant questions surround each of those approaches can be found.
- Should we specify the reference? This would remove our ability to determine the diversity of references chosen but in turn would allow us to focus more on how the different SNP methods call variant sites when working with the same reference.

5. Conclusions

The pilot PT was a useful exercise as it allowed WG4 to identify critical points for improvement prior to the full roll-out. The Wet lab part worked as anticipated and provided interesting data identifying several sequencing deviations such as contaminations and poor sequencing.

The PT indicated several quality markers which could be used and considered as future QC standards for assessing sequence quality. With the limited data submitted in the pilot PT it was, however, not possible to determine specific QC measures. This might be possible with the data from the full roll-out.

Unfortunately, it was not possible with the data submitted to the Dry lab part to conduct a participant cluster analysis assessment due to inadequate method description. This will be improved for the full roll out.

All in all, as critical point for improvement was revealed and promising data gained prior to the full roll out, the objectives for the pilot PT can be regarded as fulfilled.

6. Acknowledgement

The GMI pilot PT 2014 was supported by COMPARE, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 643476. In addition, the GMI pilot PT 2014 was supported by FDA GenomeTrakr and Microbiologics®.

Appendices

Appendix 1 Prenotification

Appendix 2 Protocol

Appendix 3 Letter to participants

Appendix 4 Descriptions of how clusters were defined as part of the Dry lab component

Reference List

1. **Aarestrup, F. M., S. E. Jorsal, P. Ahrens, N. E. Jensen, and A. Meyling**. 1997. Molecular characterization of Escherichia coli strains isolated from pigs with edema disease. J.Clin.Microbiol. **35**:20-24.

2. **Bradnam, K. R., J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W. C. Chou, J. Corbeil, F. C. Del, T. R. Docking, R. Durbin, D. Earl, S. Emrich, P. Fedotov, N. A. Fonseca, G. Ganapathy, R. A. Gibbs, S. Gnerre, E. Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J. B. Hiatt, I. Y. Ho, J. Howard, M. Hunt, S. D. Jackman, D. B. Jaffe, E. D. Jarvis, H. Jiang, S. Kazakov, P. J. Kersey, J. O. Kitzman, J. R. Knight, S. Koren, T. W. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. Maccallum, M. D. Macmanes, N. Maillet, S. Melnikov, D. Naquin, Z. Ning, T. D. Otto, B. Paten, O. S. Paulo, A. M. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F. J. Ribeiro, S. Richards, D. S. Rokhsar, J. G. Ruby, S. Scalabrin, M. C. Schatz, D. C. Schwartz, A. Sergushichev, T. Sharpe, T. I. Shaw, J. Shendure, Y. Shi, J. T. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, B. M. Vieira, J. Wang, K. C. Worley, S. Yin, S. M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, and I. F. Korf**. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience. **2**:10-12.

3. **Hendriksen, R. S., M. Mikoleit, C. Kornschober, R. L. Rickert, S. V. Duyne, C. Kjelso, H. Hasman, M. Cormican, D. Mevius, J. Threlfall, F. J. Angulo, and F. M. Aarestrup**. 2009. Emergence of Multidrug-Resistant Salmonella Concord Infections in Europe and the United States in Children Adopted From Ethiopia, 2003-2007. Pediatr.Infect.Dis.J. **28**:814-818.

4. **Kaas, R. S., P. Leekitcharoenphon, F. M. Aarestrup, and O. Lund**. 2014. Solving the problem of comparing whole bacterial genomes across different sequencing platforms. PLoS.One. **9**:e104984.

5. **Larsen, M. V., S. Cosentino, S. Rasmussen, C. Friis, H. Hasman, R. L. Marvig, L. Jelsbak, T. Sicheritz-Ponten, D. W. Ussery, F. M. Aarestrup, and O. Lund**. 2012. Multilocus sequence typing of total-genome-sequenced bacteria. J.Clin.Microbiol. **50(4)**:1355-1361.

6. **Li, H. and R. Durbin**. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. **25**:1754-1760.

7. **Moran-Gilad, J., V. Sintchenko, S. K. Pedersen, W. J. Wolfgang, J. Pettengill, E. Strain, and R. S. Hendriksen**. 2015. Proficiency testing for bacterial whole genome sequencing: an end-user survey of current capabilities, requirements and priorities. BMC.Infect.Dis. **15**:174-0902.

8. **Neidhardt, F. C.** Escherichia coli *and* Salmonella:*Cellular and Molecular Biology*. 1996. Washington DC, ASM.

9. **Zankari, E., H. Hasman, S. Cosentino, M. Vestergaard, S. Rasmussen, O. Lund, F. M. Aarestrup, and M. V. Larsen**. 2012. Identification of acquired antimicrobial resistance genes. J.Antimicrob.Chemother. **67**:2640-2644.

THE PROFICIENCY TEST (PILOT) REPORT OF THE GLOBAL MICROBIAL IDENTIFIER INITIATIVE YEAR 2014

TABLES AND FIGURES

Figure 1: Countries participating in the GMI pilot PT 2014

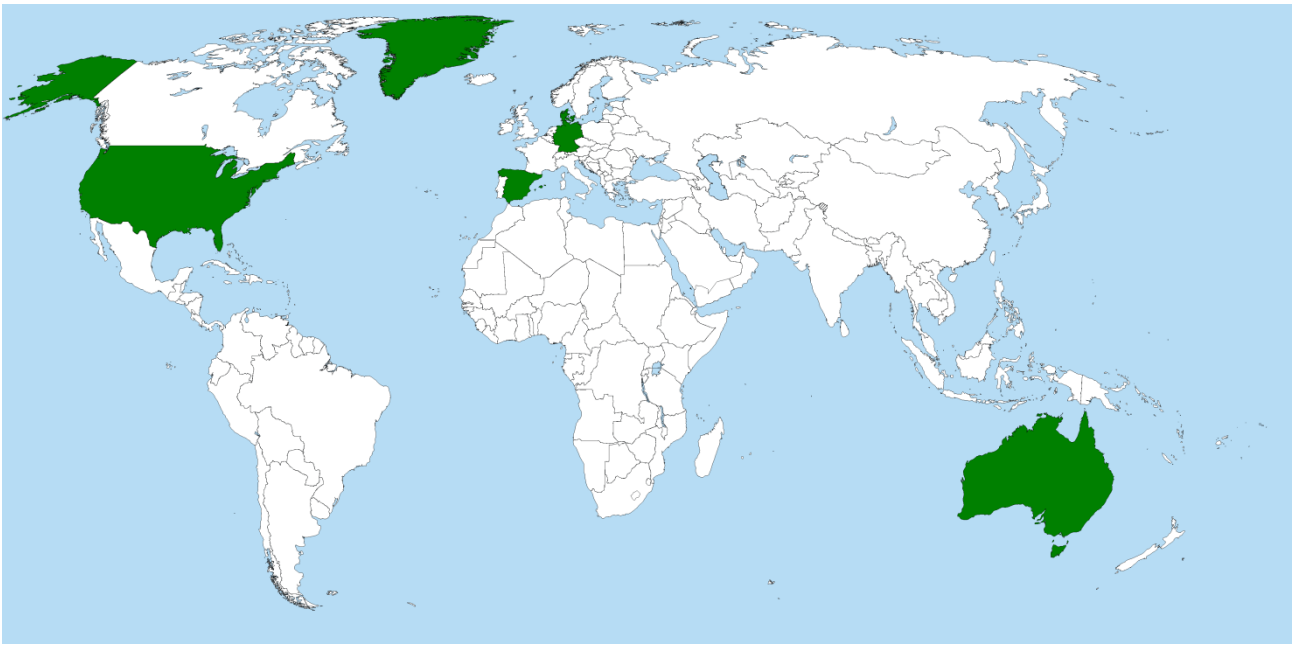


Figure 2: How the DNA concentration (ng/µl) was measured prior to library preparation for both the bacterial cultures and DNA received.

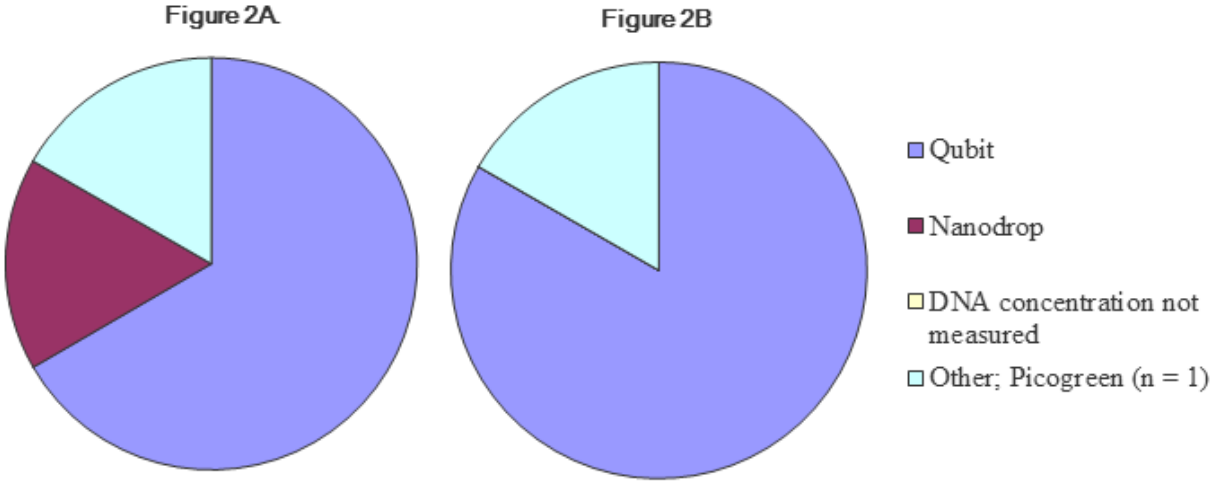


Table 1: The low and high range of DNA concentration (ng/µl) measured for both the bacterial cultures and DNA received.

	Low range (ng/µl)	High range (ng/µl)
GMI14-001-BACT (*Salmonella*)	14.1	65.2
GMI14-002-BACT (*Salmonella*)	14.3	62
GMI14-001-DNA (*Salmonella*)	0.2	113
GMI14-002-DNA (Salmonella)	0.2	128
GMI14-003-BACT (*E. coli*)	ND	ND
GMI14-004-BACT (*E. coli*)	9.64	110
GMI14-003-DNA (*E. coli*)	ND	ND
GMI14-004-DNA (*E. coli*)	0.2	80.7
GMI14-005-BACT (*S. aureus*)	13	216
GMI14-006-BACT (*S. aureus*)	9.48	127
GMI14-005-DNA (*S. aureus*)	0.2	147
GMI14-006-DNA (*S. aureus*)	0.2	87.6

Table 2: The low and high range of total DNA amount (µg) measured for both the bacterial cultures and DNA received.

	Low range (µg)	High range (µg)
GMI14-001-BACT (*Salmonella*)	2.3	6.5
GMI14-002-BACT (*Salmonella*)	1.2	6
GMI14-001-DNA (*Salmonella*)	0.02	5.65
GMI14-002-DNA (*Salmonella*)	0.02	6.4
GMI14-003-BACT (*E. coli*)	ND	ND
GMI14-004-BACT (*E. coli*)	1.9	11
GMI14-003-DNA (*E. coli*)	ND	ND
GMI14-004-DNA (*E. coli*)	0.02	7.2
GMI14-005-BACT (*S. aureus*)	1	43.2
GMI14-006-BACT (*S. aureus*)	0.7	25.4
GMI14-005-DNA (*S. aureus*)	0.02	7.9
GMI14-006-DNA (*S. aureus*)	0.02	4.38

Figure 3: Methods applied to measure the DNA quality prior to library preparation for both the bacterial cultures and DNA received.

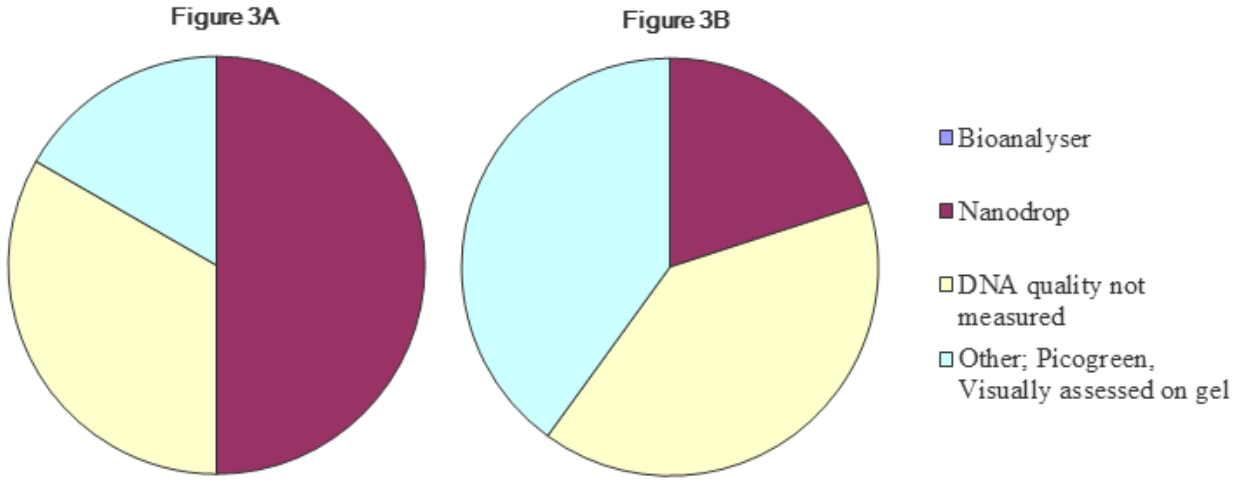


Table 3: The low and high range of the measured DNA quality (e.g. RIN or 260/280 ratio) for both the bacterial cultures and DNA received.

	Low range	High range
GMI14-001-BACT (*Salmonella*)	1.8	2.07
GMI14-002-BACT (*Salmonella*)	1.9	2.08
GMI14-001-DNA (*Salmonella*)	2.0	2.0
GMI14-002-DNA (*Salmonella*)	1.9	2.0
GMI14-003-BACT (*E. coli*)	ND	ND
GMI14-004-BACT (*E. coli*)	1.7	1.95
GMI14-003-DNA (*E. coli*)	ND	ND
GMI14-004-DNA (*E. coli*)	2.0	2.0
GMI14-005-BACT (*S. aureus*)	1.8	1.95
GMI14-006-BACT (*S. aureus*)	1.8	1.97
GMI14-005-DNA (*S. aureus*)	2.0	2.0
GMI14-006-DNA (*S. aureus*)	2.0	2.0

Table 4: The low and high range of the measured DNA quality (260/230 ratio) for both the bacterial cultures and DNA received.

	Low range	High range
GMI14-001-BACT (*Salmonella*)	2.0	2.0
GMI14-002-BACT (*Salmonella*)	1.9	1.7
GMI14-001-DNA (*Salmonella*)	0.4	0.5
GMI14-002-DNA (*Salmonella*)	0.6	0.7
GMI14-003-BACT (*E. coli*)	ND	ND
GMI14-004-BACT (*E. coli*)	1.6	1.7
GMI14-003-DNA (*E. coli*)	ND	ND
GMI14-004-DNA (*E. coli*)	0.6	0.6
GMI14-005-BACT (*S. aureus*)	1.6	2.3
GMI14-006-BACT (*S. aureus*)	1.6	2.3
GMI14-005-DNA (*S. aureus*)	0.5	0.7
GMI14-006-DNA (*S. aureus*)	0.5	0.6

Figure 4: Applied sequencing platform used in the proficiency test

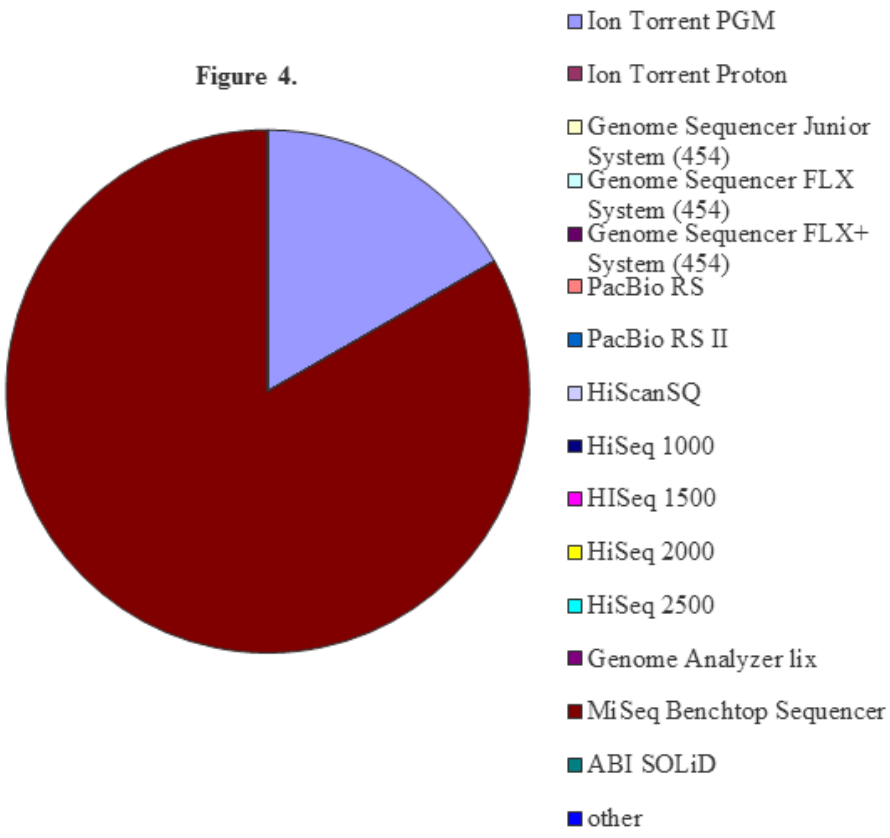


Table 5: Overview of participants' use of platforms and tools

Participant	1	2	3	4	9
Platform	Miseq	Miseq	IonTorrent	Miseq	Miseq
Read length	150	300	200	300	301
Read type	Paired-end	Paired-end	Single-end	Paired-end	Paired-end
Assembler*	Velvet	DNastar in Illumina BaseSpace	CLC Genomics Workbench 7	https://cge.cbs.dtu.dk/services/Assembler/	Velvet, https://www.ebi.ac.uk/~zerbino/velvet/, open access
Applied tools, MLST	Inhouse script	http://cge.cbs.dtu.dk/services/MLST/	MLST Databases of UoW (http://mlst.warwick.ac.uk/mlst/dbs/Senterica/); (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli) and Staphylococcus aureus MLST database (http://saureus.mlst.net/)	http://cge.cbs.dtu.dk/services/MLST	SRST2 with MLST databases from http://pubmlst.org, http://katholt.github.io/srst2/, open access
Applied tools, resistance genes	Inhouse script, ResFinder	http://cge.cbs.dtu.dk/services/ResFinder/	In-house custom pseudomolecules	http://cge.cbs.dtu.dk/services/ResFinder	SRST2 with ResFinder resistance gene database, http://katholt.github.io/srst2/, open access

* Participants assembled genomes only used for the MLST and detection antimicrobial resistance genes.

Table 6: Determined MLST and antimicrobial resistance genes in *Salmonella* **GMI14-001** for both the bacterial culture and DNA received.

	BACT										DNA									
Participant	1		2		3		4		9		1		2		3		4		9	
	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool
Expected MLST	ST19																			
Obtained MLST	19	19	19	**Unk.**	19	19	19	**1920**	**Unk.**	19	19	19	19	**Unk.**	19	19	19	19	**Unk.**	19
MLST as expected	Y	Y	Y	**N**	Y	Y	Y	**N**	**N**	Y	Y	Y	Y	**N**	Y	Y	Y	Y	**N**	Y
aroC	10	10	10	10	10	10	10	10	**19**	10	10	10	10	**10**	10	10	10	10	**19**	10
dnaN	7	7	7	**408**	7	7	7	7	**10**	7	7	7	**7**	**19**	7	7	7	7	**10**	7
hemD	12	12	12	**293**	12	12	12	12	**7**	12	12	12	12	**293**	12	12	12	12	**7**	12
hisD	9	9	9	**14**	9	9	9	**553**	**12**	9	9	9	9	**14**	9	9	9	9	**12**	9
purE	5	5	5	**444**	5	5	5	5	**9**	5	5	5	5	**176**	5	5	5	5	**9**	5
sucA	9	9	9	**109**	9	9	9	9	**5**	9	9	9	9	**103**	9	9	9	9	**5**	9
thrA	2	2	2	**104**	2	2	2	2	**9**	2	2	2	2	**104**	2	2	2	2	**9**	2
aac(6')-*Iaa*	X		X				X		X		X		X				X		X	
aac(6')-Iy		X		X		X		X		X		X		X		X		X		X

* Gene reported partly covered or a variant. MLST and resistance genes obtained by participant (own tool) and by PT-organizer (CGE tool). Results obtained by PT-organizer (CGE) shaded in grey. Deviating results indicated in bold red.

Table 7: Quality metrics related to strain GMI14-001-*Salmonella*

Participant no / sample type	Median	1_1_bac	1_1_dna	2_1_bac	2_1_dna	3_1_bac	3_1_dna	4_1_bac	4_1_dna	5_1_bac	5_1_dna	9_1_bac	9_1_dna
No. of reads	2334844	1578148	2066146	7538132	3922430	883248	1050687	3456712	2675370	2334844	1642958	5155846	3718744
Avg. read length	225.9	147.8	144.2	322.4	280.6	308.1	198.7	161.7	199.6	232.1	230.0	248.7	225.9
No. of reads that map to total reference DNA	2328486	1559868	2050075	7362866	3871448	864925	1024441	3420321	2649326	2328486	1637914	5142540	3687548
Proportion of reads that map to total reference DNA	98.95	98.84	99.22	97.67	98.70	97.93	97.50	98.95	99.03	99.73	99.69	99.74	99.16
No. of reads that map to reference chromosome	2305339	1514019	1976660	7132224	3765258	851360	992231	3304033	2579440	2305339	1583269	4960704	3566197
Proportion of reads that map to reference chromosome	96.86	97.06	96.42	96.87	97.26	98.43	96.86	96.60	97.36	99.01	96.66	96.46	96.71
No. of reads that map to reference plasmid 1	73253	46368	74407	259706	149147	13857	32579	120719	73253	24666	56573	189593	125973
Proportion of reads that map to plasmid 1	3.42	2.97	3.63	3.53	3.85	1.60	3.18	3.53	2.76	1.06	3.45	3.69	3.42
Depth of coverage, total reference DNA	106.8	46.6	59.7	479.4	219.4	53.8	41.1	111.7	106.8	109.1	76.1	258.3	168.3
Depth of coverage, total reference chromosome	106.0	46.1	58.7	473.3	217.5	54.0	40.6	110.0	106.0	110.2	75.0	254.0	165.9
Depth of coverage, total reference plasmid #1	138.5	73.0	114.2	891.3	445.6	45.4	68.9	207.8	155.6	60.9	138.5	502.0	303.0
Total size of assembly	4904497	4902142	4905044	**18070847**	**10403727**	4884330	4878995	4904497	4909253	4814319	4912476	4888969	4912967
Percentage of total size of assembly per total size of DNA	99.1	99.0	99.1	**365.0**	**210.1**	98.6	98.5	99.1	99.1	97.2	99.2	98.7	99.2
Total no. of contigs	110	122	107	**57574**	**32972**	56	65	123	110	212	96	168	87
No. of contigs > 200 bp	96	95	92	**57574**	**32972**	56	64	119	109	203	96	168	87
N50	40872	40853	40877	**150592**	**86699**	40704	40660	40872	40912	40121	40939	40743	40943
NG50	41264	41264	41264	41264	41264	41264	41264	41264	41264	41264	41264	41264	41264
Compass.pl													
Coverage (by summing "island" lengths)	0.3669	0.4785	0.5180	0.0386	0.0309	0.5246	0.5263	0.3669	0.3425	0.2091	0.3895	0.2489	0.4702
Validity	0.3705	0.4835	0.5230	0.0107	0.0149	0.5319	0.5342	0.3705	0.3456	0.2151	0.3926	0.2522	0.4740
Multiplicity	1.0003	1.0003	1.0003	1.0148	1.0094	1.0001	1.0000	1.0004	1.0004	1.0002	1.0002	1.0007	1.0002
Parsimony (Multiplicity / Validity)	2.5475	2.0691	1.9126	94.4762	67.9015	1.8803	1.8721	2.6998	2.8948	4.6497	2.5475	3.9675	2.1101

REFERENCES		**Size, total DNA**	**Size, chromosome**	**Size, plasmid #1**
CFSAN18746 Salmonella	GMI14-001	4951491	4857558	93933

Shading indicates the lowest value for each parameter.

Shading indicates the highest value for each parameter.

All data assembled using the CGE assembler. Deviating results indicated in bold red.

Table 8: Determined MLST and antimicrobial resistance genes in *Salmonella* **GMI14-002** for both the bacterial culture and DNA received.

	BACT										DNA									
Participant	1		2		3		4		9		1		2		3		4		9	
	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool
Expected MLST	ST599																			
Obtained MLST	599	599	599	599	599	599	599	599	**Unk.**	599	599	599	599	599	599	599	599	599	**Unk.**	599
MLST as expected	Y	Y	Y	Y	Y	Y	Y	Y	**N**	Y	Y	Y	Y	Y	Y	Y	Y	Y	**N**	Y
aroC	14	14	14	14	14	14	14	14	**599**	14	14	14	14	14	14	14	14	14	**599**	14
dnaN	7	7	7	7	7	7	7	7	**14**	7	7	7	7	7	7	7	7	7	**14**	7
hemD	3	3	3	3	3	3	3	3	**7**	3	3	3	3	3	3	3	3	3	**7**	3
hisD	12	12	12	12	12	12	12	12	**3**	12	12	12	12	12	12	12	12	12	**3**	12
purE	6	6	6	6	6	6	6	6	**12**	6	6	6	6	6	6	6	6	6	**12**	6
sucA	19	19	19	19	19	19	19	19	**6**	19	19	19	19	19	19	19	19	19	**6**	19
thrA	12	12	12	12	12	12	12	12	**19**	12	12	12	12	12	12	12	12	12	**19**	12
bla$_{\text{CTX-M-15}}$	X	X	X	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X
bla$_{\text{SHV-12}}$	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
bla$_{\text{TEM-1b}}$	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
*cat*A2	**X***						**X**		**X**		**X***						**X**		**X**	
*dfr*A18	X*	X		X		X	X	X	X	X	X*	X		X		X	X	X	X	X
*dfr*A19					**X**										**X**					
ere(A)	X*	X		X	X	X	X	X	X	X	X*	X		X	X	X	X	X	X	X
*flo*R	X*	X	X	X		X	X	X		X	X*	X	X	X		X	X	X		X
*pbr*A-*pbr*B					**X**										**X**					
qacEdelta1					**X**										**X**					
*Qnr*B2					**X**				**X**						**X**				**X**	
*Qnr*B49	X*	X		X		X	X	X		X	X*	X		X		X	X	X		X
*str*A	X	X	X	X		X	X	X	X		X	X	X			X	X	X	X	X
*str*B	X	X		X		X	X	X	X	X	X	X		X		X	X	X	X	X
*sul*1	X	X	X	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X
*sul*2	X	X	X	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X
tet(A)	X*	X	X	X	X	X	X	X		X	X*	X	X	X	X	X	X	X		X
tet(D)	X	X		X		X	X	X	X	X	X	X		X		X	X	X	X	X
aac(3)-IIb							**X**										**X**			
aac(6')-IIc	X	X		X		X	X	X	X	X		X		X		X	X	X	X	X
aac(6')-Iy	X*	X	X	X		X	X	X	X	X	X*	X	X	X		X	X	X	X	X
*aac*3					**X**										**X**					
*aac*A27					**X**										**X**					
arr					**X**										**X**					

* Gene reported partly covered or a variant. MLST and resistance genes obtained by participant (own tool) and by PT-organizer (CGE tool). Results obtained by PT-organizer (CGE) shaded in grey. Deviating results indicated in bold red.

Table 9: Quality metrics related to strain GMI14-002-*Salmonella*

Participant no / sample type	Median	1_2_bac	1_2_dna	2_2_bac	2_2_dna	3_2_bac	3_2_dna	4_2_bac	4_2_dna	5_2_bac	5_2_dna	9_2_bac	9_2_dna
No. of reads	1900294	1881906	1900294	3345320	4480666	1088563	1335123	1475002	2931292	2225058	1246870	2620206	3160322
Avg. read length	220.0	145.9	139.5	323.0	252.3	313.1	187.2	145.5	196.4	229.9	220.0	279.7	238.9
No. of reads that map to total reference DNA	1873000	1832839	1873000	3168961	4419037	1065848	1302305	1448703	2907041	2219823	1241530	2613839	3142859
Proportion of reads that map to total reference DNA	98.56	97.39	98.56	94.73	98.62	97.91	97.54	98.22	99.17	99.76	99.57	99.76	99.45
No. of reads that map to reference chromosome	1572970	1572970	1529203	2664573	3764174	944841	1080184	1178540	2451004	1968566	1016429	2224595	2572503
Proportion of reads that map to reference chromosome	84.08	85.82	81.64	84.08	85.18	88.65	82.94	81.35	84.31	88.68	81.87	85.11	81.85
No. of reads that map to reference plasmid 1	245039	218435	310564	478060	768487	115302	202963	245039	402003	242540	204991	365528	519129
Proportion of reads that map to plasmid 1	15.09	11.92	16.58	15.09	17.39	10.82	15.58	16.91	13.83	10.93	16.51	13.98	16.52
No. of reads that map to reference plasmid 2	29318	43946	36149	52903	83356	7157	24468	29056	62268	13368	23737	29318	62035
Proportion of reads that map to plasmid 2	1.89	2.40	1.93	1.67	1.89	0.67	1.88	2.01	2.14	0.06	1.91	1.12	1.97
Depth of coverage, total reference DNA	63.2	50.7	49.5	194.0	211.2	63.2	46.2	39.9	108.2	96.7	51.8	138.5	142.3
Depth of coverage, total reference chromosome	62.4	48.4	45.0	181.6	200.3	62.4	42.6	36.2	101.5	95.5	47.2	131.2	129.6
Depth of coverage, total reference plasmid #1	81.2	71.7	97.5	347.5	436.3	81.2	85.5	80.2	177.7	125.5	101.5	230.0	279.0
Depth of coverage, total reference plasmid #2	56.5	69.4	54.6	185.0	227.7	24.3	49.6	45.8	132.4	33.3	56.5	88.8	160.4
Total size of assembly	5187715	5191393	5187715	5226456	5216534	5169263	5168642	5182363	5190490	5074396	5189716	5182958	5191563
Percentage of total size of assembly per total size of DNA	98.3	98.4	98.3	99.0	98.8	97.9	97.9	98.2	98.3	96.1	98.3	98.2	98.4
Total no. of contigs	150	150	186	122	139	77	77	359	171	187	192	147	171
No. of contigs > 200 bp	113	113	133	122	139	77	76	298	159	167	168	147	148
N50	43232	43263	43232	43555	43473	43079	43074	43188	43256	42288	43249	43193	43265
NG50	43982	43982	43982	43982	43982	43982	43982	43982	43982	43982	43982	43982	43982
Compass.pl													
Coverage (by summing "island" lengths)	0.3733	0.4305	0.3999	0.6110	0.4061	0.3924	0.3634	0.2178	0.3083	0.2698	0.3226	0.4147	0.3733
Validity	0.3561	0.4109	0.3819	0.5792	0.3856	0.3760	0.3483	0.2082	0.2942	0.2635	0.3079	0.3963	0.3561
Multiplicity	1.0003	1.0007	1.0004	1.0006	1.0003	1.0003	1.0003	1.0003	1.0004	1.0009	1.0002	1.0003	1.0001
Parsimony (Multiplicity / Validity)	2.660	2.435	2.620	1.727	2.594	2.660	2.872	4.805	3.401	3.798	3.249	2.524	2.809
REFERENCES		**Size, total DNA**		**Size, chromosome**		**Size, plasmid #1**		**Size, plasmid #2**					
CFSAN018747 *Salmonella*	GMI14-002	5277621		4740838		444417		92366					

Shading indicates the lowest value for each parameter.

Shading indicates the highest value for each parameter.

All data assembled using the CGE assembler. Deviating results indicated in bold red.

Table 10: Determined MLST and antimicrobial resistance genes in *E. coli* **GMI14-004** for both the bacterial culture and DNA received.

Participant	BACT*										DNA*									
	1		2		3		4		9		1		2		3		4		9	
	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool
Expected MLST	adK 233, fumC 2, gyrB 29, icD 167, mdH 4, purA 16, recA 4																			
Obtained MLST	Unk.	Unk.	Unk.	Unk.	Unk.	Unk.	Unk.	Unk.	Unk.	Unk.	Unk.	Unk.	Unk.	Unk.	Unk.	Unk.	Unk.	Unk.	Unk.	Unk.
MLST as expected	-	Y	-	**N**	**N**	**N**	**N**	Y	**N**	Y	-	Y	-	**N**	**N**	**N**	**N**	Y	**N**	Y
adK	-	233	-	233	233	233	**10**	233	233	233	-	233	-	**138**	233	233	**10**	233	233	233
fumC	-	2	-	2	2	2	**7**	2	**233**	2	-	2	-	**486**	2	2	**7**	2	**233**	2
gyrB	-	29	-	29	29	29	**12**	29	**2**	29	-	29	-	**181**	29	29	**12**	29	**2**	29
icD	-	167	-	167	167	167	**9**	167	**29**	167	-	167	-	167	167	167	**9**	167	**29**	167
mdH	-	4	-	4	4	4	**5**	4	**167**	4	-	4	-	**374**	4	4	**5**	4	**167**	4
purA	-	16	-	**12**	16	16	**9**	16	**4**	16	-	16	-	**1**	16	16	**9**	16	**4**	16
recA	-	4	-	4	**99**	**215**	**2**	4	**16**	4	-	4	-	**189**	**99**	**215**	**2**	4	**16**	4
aac(6')-IIc				**X**									**X**	**X**						
aac(6')-Iy														**X**						
ere(A)													**X**	**X**						
*dfr*A18													**X**	**X**						
*Qnr*B49				**X**									**X**	**X**						
*str*A				**X**									**X**	**X**						
*str*B													**X**	**X**						
*sul*1				**X**									**X**	**X**						
*sul*2				**X**									**X**	**X**						
floR														**X**						
tet(A)														**X**						
tet(D)														**X**						
$bla_{\text{TEM-1b}}$														**X**						

MLST and resistance genes obtained by participant (own tool) and by PT-organizer (CGE tool). Results obtained by PT-organizer (CGE) shaded in grey. Deviating results indicated in bold red. *Approximately 15% of the sequence reads does not map to the reference DNA suggesting some contamination.

Table 11: Quality metrics related to strain GMI14-004-*E.coli*

Participant no / sample type	Median	1_4_bac	1_4_dna	2_4_bac	2_4_dna	3_4_bac	3_4_dna	4_4_bac	4_4_dna	5_4_bac	5_4_dna	9_4_bac	9_4_dna
No. of reads	2053210	2053210	2053210	7551992	5069204	1106824	968124	2029982	2562486	1821444	159360	4044036	3744542
avg. read length	146.42	146.4	146.4	321.9	300.2	304.6	184.2	144.5	179.6	231.3	117.2	263.4	231.3
No. of reads that map to total reference DNA	1849332	1849332	1849332	6695516	4281238	1016206	781966	1675927	2370359	1728947	129474	3849816	3124724
Proportion of reads that map to total reference DNA	88.7	90.1	90.1	88.7	84.5	91.8	80.8	82.6	92.5	94.9	81.2	95.2	83.4
No. of reads that map to reference chromosome	1725471	1725471	1725471	6310855	4078215	1001455	736661	1577019	2315755	1696390	121796	3751380	2941166
Proportion of reads that map to reference chromosome	94.2	93.3	93.3	94.3	95.3	98.5	94.2	94.1	97.7	98.1	94.1	97.4	94.1
No. of reads that map to reference plasmid 1	36621	54254	54254	155367	116081	6541	18260	39847	23562	16505	3029	36621	72892
Proportion of reads that map to plasmid 1	2.33	2.93	2.93	2.32	2.71	0.64	2.34	2.38	0.99	0.95	2.34	0.95	2.33
No. of reads that map to reference plasmid 2	22643	29445	29445	134065	72093	3799	13268	30037	15280	10093	2372	22643	52222
Proportion of reads that map to plasmid 2	1.59	1.59	1.59	2.00	1.68	0.37	1.70	1.79	0.64	0.58	1.83	0.59	1.67
No. of reads that map to reference plasmid 3	62260	72962	72962	279329	171233	14863	25858	62260	46543	25156	4516	103866	111694
Proportion of reads that map to plasmid 3	3.48	3.95	3.95	4.17	4.00	1.46	3.31	3.71	1.96	1.45	3.49	2.70	3.57
Depth of coverage. total reference DNA	59.06	51.66	51.66	411.21	245.18	59.06	27.47	46.2	81.21	76.29	**2.9**	193.48	137.91
Depth of coverage. total reference chromosome	60.16	49.82	49.82	400.62	241.41	60.16	26.75	44.94	82.01	77.37	**2.82**	194.88	134.18
Depth of coverage. total reference plasmid #1	37.6516	150.1	150.1	945.2	658.5	37.7	63.5	108.8	80.0	72.1	**6.7**	182.3	318.7
Depth of coverage. total reference plasmid #2	23.76981	88.6	88.6	886.5	444.5	23.8	50.2	89.1	56.4	48.0	**5.7**	122.5	248.2
Depth of coverage. total reference plasmid #3	65.66213	154.9	154.9	1304.2	745.5	65.7	69.1	130.5	121.2	84.4	**7.7**	396.8	374.8
Total size of assembly	5330032	5397127	5397011	**19331077**	**13180319**	5330032	5293359	5367412	5351085	5165154	**1278221**	5401414	5456878
***Percentage of total size of assembly per total size of DNA**	**102.4**	**103.0**	**103.0**	**368.8**	**251.5**	**101.7**	**101.0**	**102.4**	**102.1**	**98.5**	**24.4**	**103.0**	**104.1**
Total no. of contigs	363	409	409	**62630**	**12311**	212	243	728	433	363	353	315	334
No. of contigs > 200 bp	353	387	387	**62630**	**12311**	212	242	593	391	353	352	315	334
N50	44730	44978	44977	**161094**	**109837**	44418	44113	44730	44594	43044	43044	45013	45475
NG50	43682	43682	43682	43682	43682	43682	43682	43682	43682	43682	10653	43682	43682
Compass.pl													
Coverage (by summing "island" lengths)	0,001123	0.0019	0.0008	0.0082	0.0591	0.0005	0.0008	0.0013	0.0011	0.0013	0.0010	0.0008	0.0015
Validity	0,001039	0.0017	0.0008	0.0022	0.0223	0.0005	0.0008	0.0012	0.0010	0.0013	0.0037	0.0007	0.0014
Multiplicity	1.0000	1.0000	1.0469	1.0236	1.0039	1.0000	1.0000	1.0027	1.0000	1.0142	1.0000	1.0000	1.0173
Parsimony (Multiplicity / Validity)	794,272	580.5	1389.6	474.5	45.0	2066.7	1300.9	813.9	962.4	794.3	268.7	1390.0	714.3
REFERENCES		**Size. total DNA**		**Size. chromosome**		**Size. plasmid #1**		**Size. plasmid #2**		**Size. plasmid #3**			
CFSAN018749 E. coli	GMI14-004	5241609		5071057		52918		48684		68950			

Shading indicates the lowest value for each parameter.

Shading indicates the highest value for each parameter.

All data assembled using the CGE assembler. Deviating results indicated in bold red. *Approximately 15% of the sequence reads does not map to the reference DNA suggesting some contamination.

Table 12: Determined MLST and antimicrobial resistance genes in *S. aureus* **GMI14-005** for both the bacterial culture and DNA received. MLST and resistance genes obtained by participant (own data) and by PT-organizer (CGE).

	BACT										DNA									
Participant	1		2		3		4		9		1		2		3		4		9	
	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool
Expected MLST	ST433																			
Obtained MLST	433	433	433	**Unk**	433	433	433	433	**Unk**	433	433	433	433	433	433	433	433	433	**Unk**	433
MLST as expected	Y	Y	Y	**N**	Y	Y	Y	Y	**N**	Y	Y	Y	Y	Y	Y	Y	Y	Y	**N**	Y
arcC	2	2	2	**77**	2	2	2	2	**433**	2	2	2	2	2	2	2	2	2	**9**	2
aroE	2	2	2	**148**	2	2	2	2	2	2	2	2	2	2	2	2	2	2	**3**	2
glpF	2	2	2	**151**	2	2	2	2	2	2	2	2	2	2	2	2	2	2	**3**	2
Gmk	2	2	2	**109**	2	2	2	2	2	2	2	2	2	2	2	2	2	2	**1**	2
Pta	6	6	6	**112**	6	6	6	6	**2**	6	6	6	6	6	6	6	6	6	**1**	6
Tpi	3	3	3	**123**	3	3	3	3	**6**	3	3	3	3	3	3	3	3	3	**1**	3
yqiL	72	72	72	**277**	72	72	72	72	**3**	72	72	72	72	72	72	72	72	72	**1**	72
aac(3)-Ik									**X**										**X**	
blaZ	X	X	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
*czr*C					**X**										**X**					
*fus*A5									**X**										**X**	
*mec*A(10)	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
*nor*A							**X**										**X**			
*str*2		X	X			X	X	X	X	X		X				X	X	X	X	X
tet(38)	X	X	X			X	X	X	X	X	X	X	X			X	X	X	X	X
vga(A)					**X**		**X**								**X**		**X**			

Results obtained by PT-organizer (CGE) shaded in grey.
MLST and resistance genes obtained by participant (own tool) and by PT-organizer (CGE tool).
Results obtained by PT-organizer (CGE) shaded in grey. Deviating results indicated in bold red.

Table 13: Quality metrics related to strain GMI14-005-*S.aureus*

Participant no / sample type	Median	1_5_bac	1_5_dna	2_5_bac	2_5_dna	3_5_bac	3_5_dna	4_5_bac	4_5_dna	5_5_bac	5_5_dna	9_5_bac	9_5_dna
No. of reads	2380956	2380956	1573038	3008880	4904674	1108555	1053731	5844236	7243770	1422214	1988694	4341732	4201120
Avg. read length	222.17	143.0	145.6	293.3	306.3	301.9	174.0	192.9	170.9	222.2	225.1	253.8	271.3
No. of reads that map to total reference DNA	2273852	2273852	1529393	2781882	4770766	1076971	1006197	5731966	6933101	1411053	1964423	4296096	4139215
Proportion of reads that map to total reference DNA	97.22	95.50	97.20	92.50	97.30	97.20	95.50	98.10	95.70	99.20	98.80	98.90	0.9850
No. of reads that map to reference chromosome	2107796	2107796	1431222	2573148	4677089	1056727	980277	5539930	6637614	1371752	1898583	4153769	3996512
Proportion of reads that map to reference chromosome	96.64	92.70	93.60	92.50	98.00	98.10	97.40	96.60	95.70	97.20	96.60	96.70	0.9660
No. of reads that map to reference plasmid 1	118408	167334	98902	231197	118408	20323	26104	199273	305807	40674	68342	147529	147464
Proportion of reads that map to plasmid 1	3.47	7.36	6.47	8.31	2.48	1.89	2.59	3.48	4.41	2.88	3.48	3.43	0.0356
Depth of coverage. total reference DNA	151.5	111.4	76.3	279.6	500.8	111.4	60.0	379.0	405.9	107.4	151.5	373.6	384.8
Depth of coverage. total reference chromosome	146.6	103.4	71.5	259.0	491.7	109.5	58.5	366.8	389.2	104.6	146.7	361.8	372.1
Depth of coverage. total reference plasmid #1	5441.2	5441.3	3275.4	15421.6	8248.7	1395.2	1032.7	8743.8	11882.8	2055.2	3498.2	8514.2	9097.8
Total size of assembly	2857685	2846735	2857685	**3938151**	**15282564**	2850812	2850496	2881232	2861935	2732285	2762642	2866664	2881158
Percentage of total size of assembly per total size of DNA	97.9	97.6	97.9	**135.0**	**523.7**	97.7	97.7	98.7	98.1	93.6	94.7	98.2	98.7
Total no. of contigs	144	106	124	**3907**	**51020**	64	65	144	170	212	182	176	118
No. of contigs > 200 bp	143	99	100	**3907**	**51020**	64	64	143	165	206	182	176	118
N50	23816	23724	23816	**32819**	**127356**	23758	23756	24012	23851	22771	23024	23890	24011
NG50	24319	24319	24319	24319	24319	24319	24319	24319	24319	24319	24319	24319	24319
Compass.pl													
Coverage (by summing "island" lengths)	0.00002323	0.00002323	0.00002323	0.00005029	0.00177341	0.00000020	0.00001171	0.00002323	0.00002323	0.00002323	0.00002323	0.00002323	0.00002323
Validity	0.00005935	0.00004005	0.00005984	0.00006297	0.00057451	NA	0.00003999	0.00005935	0.00003983	0.00006259	0.00004126	0.00003977	0.00005935
Multiplicity	1.500	1.0000	1.5000	1.0000	1.0000	NA	2.0000	1.5000	1.0000	1.5000	1.0000	1.0000	1.5000
Parsimony (Multiplicity / Validity)	25067.4	24971.4	25067.4	15879.6	1740.6	NA	50008.7	25274.0	25104.7	23967.4	24233.7	25146.2	25273.3

REFERENCES		**Size. total DNA**	**Size. chromosome**	**Size. plasmid #1**
CFSAN018750 *S. aureus*	GMI14-005	2918141	2913744	4397

cates the lowest value for each parameter.

cates the highest value for each parameter.

All data assembled using the CGE assembler. Deviating results indicated in bold red.

Figure 5: Phylogenetic SNP analysis of strain GMI14-005-*S.aureus*

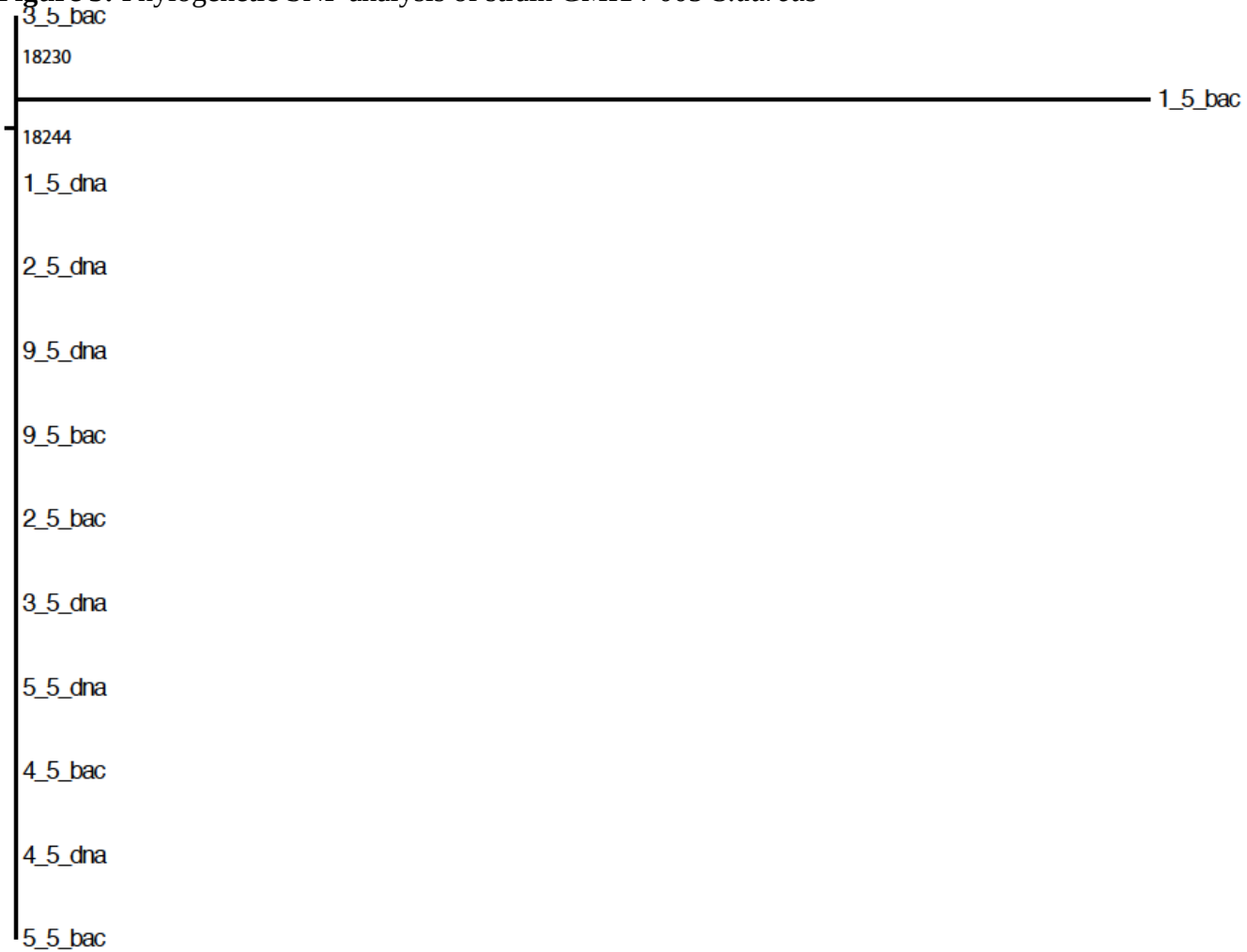


Table 14: Determined MLST and antimicrobial resistance genes in *S. aureus* **GMI14-006** for both the bacterial culture and DNA received. MLST and resistance genes obtained by participant (own data) and by PT-organizer (CGE).

	BACT										DNA									
Participant	1		2		3		4		9		1		2		3		4		9	
	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool	Own tool	CGE tool
Expected MLST	ST9																			
Obtained MLST	9	9	9	**Unk.**	9	9	9	9	**Unk.**	9	9	9	9	**Unk.**	9	9	9	9	Unk.	9
MLST as expected	Y	Y	Y	**N**	Y	Y	Y	Y	**N**	Y	Y	Y	Y	**N**	Y	Y	Y	Y	N	Y
arcC	3	3	3	**90**	3	3	3	3	**9**	3	3	3	3	**90**	3	3	3	3	9	3
aroE	3	3	3	**109**	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
glpF	1	1	1	**205**	1	1	1	1	**3**	1	1	1	1	1	1	1	1	1	**3**	1
Gmk	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Pta	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tpi	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
yqiL	10	10	10	**277**	10	10	10	10	**1**	10	10	10	10	**277**	10	10	10	10	**1**	10
aac(3)-Ik									**X**										**X**	
blaZ			**X**										**X**							
fosB					**X**										**X**					
*fusA*6									**X**										**X**	
mecA			**X**										**X**							
norA							**X**										**X**			
tet(38)	X	X	X	X		X	X	X	X	X	X	X	X			X	X	X	X	X

Results obtained by PT-organizer (CGE) shaded in grey.
MLST and resistance genes obtained by participant (own tool) and by PT-organizer (CGE tool).
Results obtained by PT-organizer (CGE) shaded in grey. Deviating results indicated in bold red.

Table 15: Quality metrics related to strain GMI14-006-*S.aureus*

Participant no / sample type	Median	1_6_bac	1_6_dna	2_6_bac	2_6_dna	3_6_bac	3_6_dna	4_6_bac	4_6_dna	5_6_bac	5_6_dna	9_6_bac	9_6_dna
No. of reads	2532644	2202384	2532644	3068108	3540796	995773	1030584	5890072	5577444	1525306	1719636	3602104	4799982
Avg. read length	220.9	146.9	147.1	307.7	306.8	299.7	191.6	186.1	175.2	220.9	222.2	270.9	268.4
No. of reads that map to total reference DNA	2528015	2193863	2528015	2736537	3470566	961288	997951	5856834	5546314	1521060	1714324	3597045	4791410
Proportion of reads that map to total reference DNA	99.44	99.60	99.80	89.20	98.00	96.50	96.80	99.40	99.40	99.70	99.70	99.90	99.80
No. of reads that map to reference chromosome	2528015	2193863	2528015	2736537	3470566	961288	997951	5856834	5546314	1521060	1714324	3597045	4791410
Proportion of reads that map to reference chromosome	1	1	1	1	1	1	1	1	1	1	1	1	1
Depth of coverage. total reference DNA	140.1	118.5	136.7	309.7	391.6	105.9	70.3	400.8	357.3	123.5	140.1	358.3	473.0
Depth of coverage. total reference chromosome	140.1	118.5	136.7	309.7	391.6	105.9	70.3	400.8	357.3	123.5	140.1	358.3	473.0
Total size of assembly	2665839	2656448	2642838	**5934985**	**11891802**	2665814	2665839	2684724	2669724	2539475	2546573	2676813	2680210
Percentage of total size of assembly per total size of DNA	98.0	97.7	97.2	**218.2**	**437.3**	98.0	98.0	98.7	98.2	93.4	93.6	98.4	98.6
Total no. of contigs	104	41	46	**6677**	**39434**	39	41	104	106	169	184	106	87
No. of contigs > 200 bp	103	40	42	**6677**	**39434**	39	40	103	102	168	183	106	87
N50	22217	22139	22025	**49460**	**99100**	22217	22217	22374	22249	21164	21223	22308	22337
NG50	22663	22663	22663	22663	22663	22663	22663	22663	22663	22663	22663	22663	22663
Compass.pl													
Coverage (by summing "island" lengths)	0.0000232	0.0000303	0.0000188	0.0000580	0.0022420	0.0000188	0.0000303	0.0000232	0.0000232	0.0000232	0.0000347	0.0000303	0.0000002
Validity	0.0000559	0.0000775	0.0000348	0.0000482	0.0009334	0.0000345	0.0000559	0.0000637	0.0000427	0.0000449	0.0000671	0.0000770	NA
Multiplicity	1.0	1.4	1.0	1.0	1.0	1.0	1.0	1.5	1.0	1.0	1.0	1.4	NA
Parsimony (Multiplicity / Validity)	20751.7	17828.5	28726.5	20751.7	1071.3	28976.2	17891.5	23550.2	23418.6	22276.1	14892.2	17965.2	NA
REFERENCES			**Size. total DNA**		**Size. chromosome**								
CFSAN018751 *S. aureus*	GMI14-006	2719423		2719423									

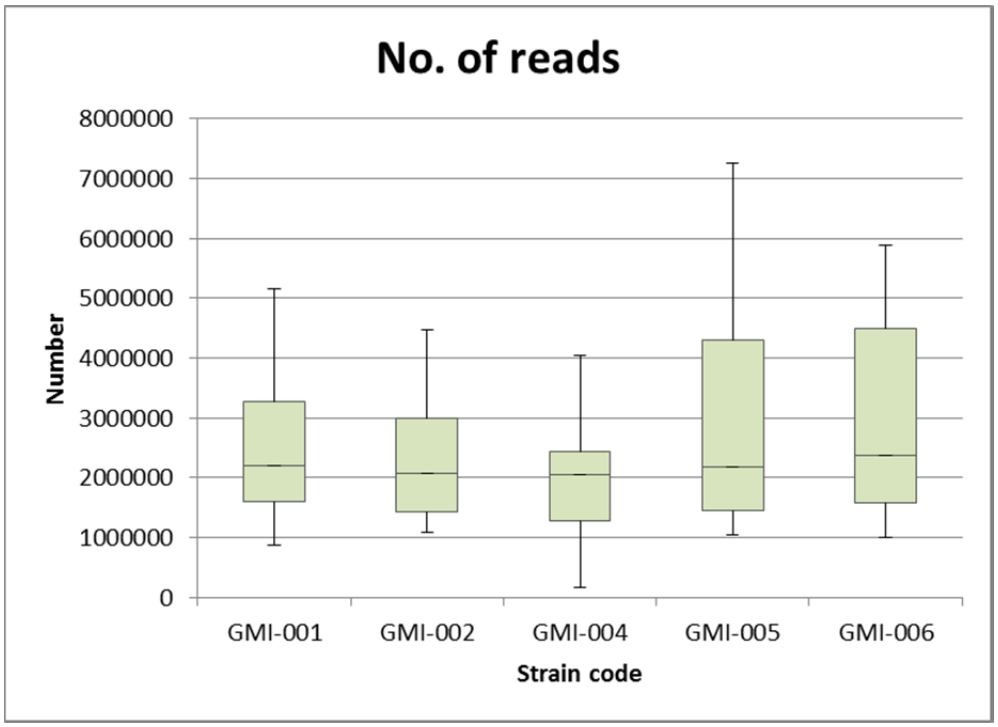
Shading indicates the lowest value for each parameter.

Shading indicates the highest value for each parameter.

All data assembled using the CGE assembler. Deviating results indicated in bold red.

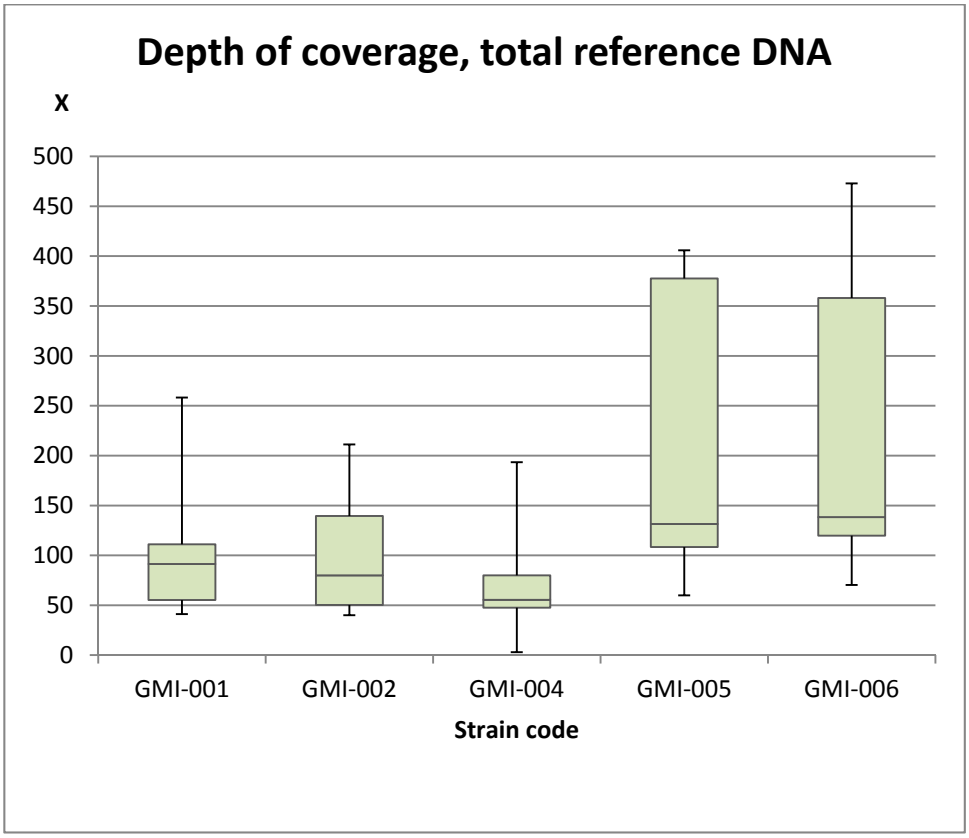
Figure 6: Phylogenetic SNP analysis of strain GMI14-006-*S.aureus*

Figure 6: Number of reads per strain and per participant



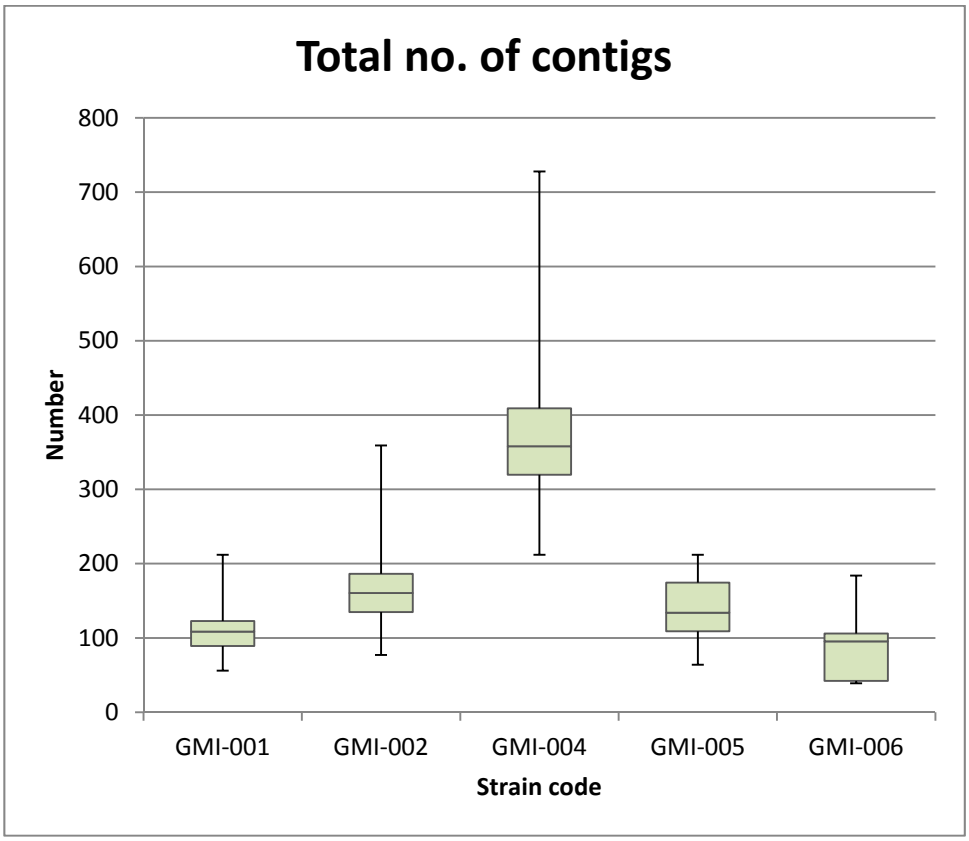
Results for participant 2 omitted for both sample types of strain 1, 4, 5, and 6. The whiskers represent minimum and maximum values (range) and the box represent the Q1, Median, and Q3, respectively.

Figure 7: Depth (x) of coverage per strain and per participant



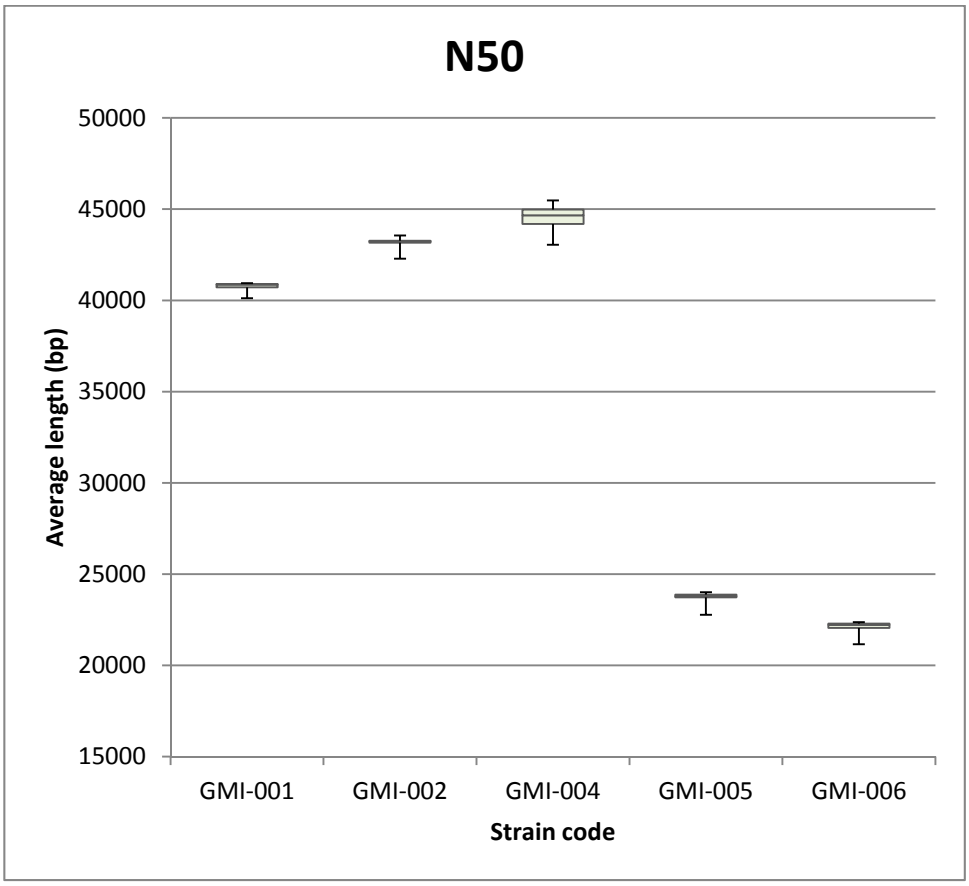
Results for participant 2 omitted for both sample types of strain 1, 4, 5, and 6. The whiskers represent minimum and maximum values (range) and the box represent the Q1, Median, and Q3, respectively.

Figure 8: Total number of contigs per strain and per participant



Results for participant 2 omitted for both sample types of strain 1, 4, 5, and 6. The whiskers represent minimum and maximum values (range) and the box represent the Q1, Median, and Q3, respectively.

Figure 9: N50 (average length (bp) of sequences) per strain and per participant



Results for participant 2 omitted for both sample types of strain 1, 4, 5, and 6. The whiskers represent minimum and maximum values (range) and the box represent the Q1, Median, and Q3, respectively.

Table 16: Summary of assemblers used by participants of the dry-lab-component.

Assembler	N
CLC Genomics Workbench 7	1
n/a	2
NGOPT (http://sourceforge.net/projects/ngopt/)	2
Velvet (https://www.ebi.ac.uk/~zerbino/velvet/)	3

Table 17: Diversity of reference genomes within each taxonomic group used by the participating laboratories.

Reference	N
Staphylococcus aureus M1	1
Staphylococcus aureus NCTC 8325	1
Staphylococcus aureus MSSA476	1
Staphylococcus aureus USA300	1
Staphylococcus aureus MSSA476	2
Staphylococcus aureus M013	1
Staphylococcus aureus MW2	1
Escherichia coli O111:H- str 11128	6
Escherichia coli ATCC BAA-2209	1
Escherichia coli O157:H7	1
Salmonella Typhimurium SL1344	1
Salmonella Typhimurium SL1344	1
Salmonella Typhimurium 14028S	1
Salmonella Typhimurium LT2	5

Table 18: Number of individuals (*N*) and length of the matrices constructed by each participant for each of the three Dry lab datasets.

	E. coli		*S. aureus*		*S.* Typhimurium	
Lab	*N*	Length	*N*	Length	*N*	Length
1	23	18811	25	1590	21	11481
2	22	58751	24	22723	20	15304
3	22	3806	24	11378	19	8662
4	22	36611	24	1439	20	18181
5	23	13196	25	1775	21	10567
6	18	150447	25	5903	22	19881
7	23	26875	25	4373	21	15586

Figure 10: Scatterplot illustrating the differences in the size of the SNP matrix (x-axis) and average pairwise distances among samples (y-axis) for each of the three datasets (coloring of points) across the different participants (individual points)

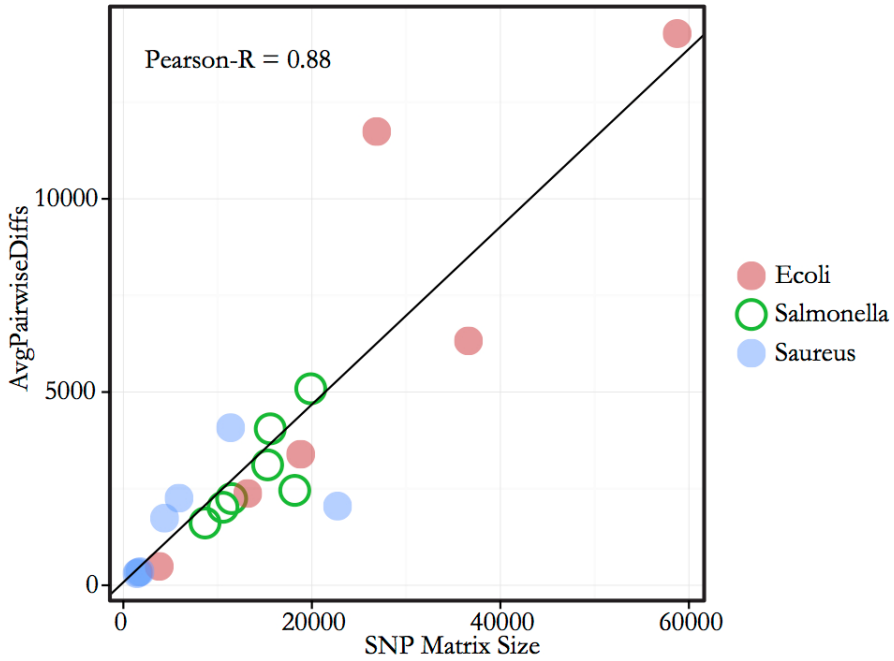


Figure 11: Heatmaps illustrating the number of pairwise SNP differences within each of the three datasets for each participating laboratory. The y-axis are pairwise comaprisons and sorted to be the same across the participants. The figure illustrates that despite differences in protocols and number of SNPs that the participants often detected similar patterns of differentiation but that was not always the case (e.g., Lab3 S. aureus detected large pairwise differences among all samples, which was not the case for the other participants).

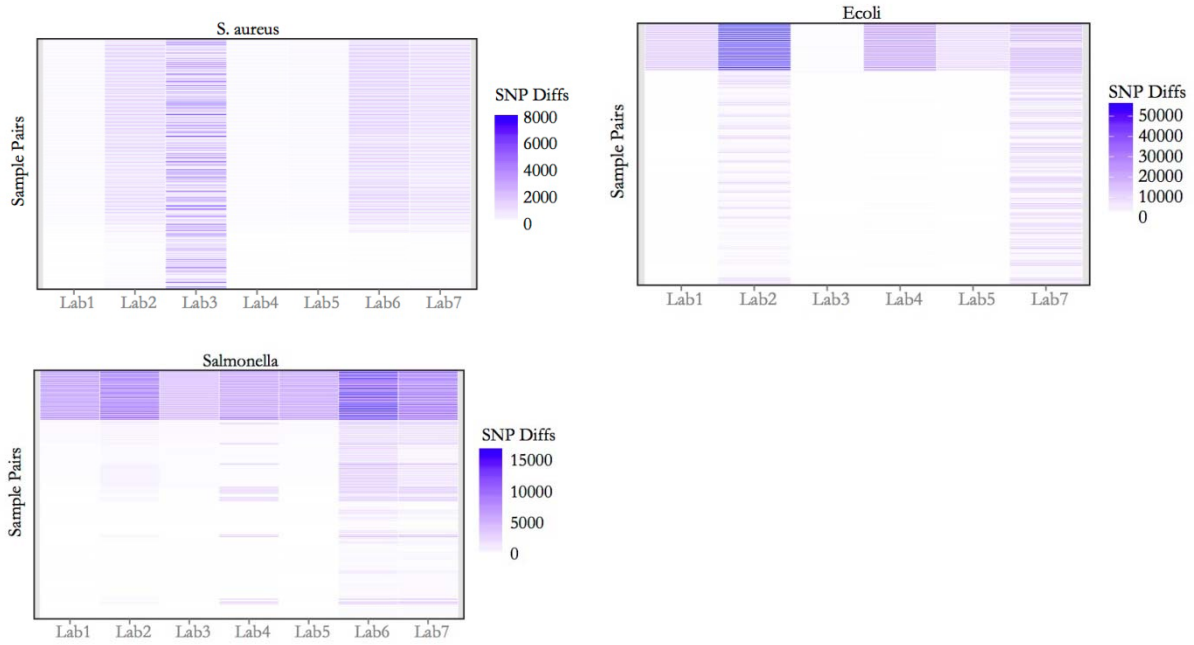


Table 19: The number of clusters within each dataset identified by the participants and the method described to identify those clusters.

Lab	*S.* Typhimurium	*E. coli*	*S. aureus*	Method
1	2	2	6	CTree (http://www.phylogenetictrees.com/ctree.php)
2	2	n/a	2	Relative SNP differences
3	Ill-formatted result from SurveyMonkey	4	2	
4	3	2	3	Not described
5	3	2	2	Not described
6	3	3	2	Not described
7	3	4	3	SNP differences
8	6	5	4	Based on phylogenetic analysis, MLST and resistance genes

Denmark, June 2014

Prenotification for GMI proficiency test pilot, 2014

GMI is a global, visionary taskforce of scientists and other stakeholders who shares an aim of making novel genomic technologies and informatics tools available for improved global patient diagnostics, surveillance and research, by developing needs- and end-user-based data exchange and analysis tools for characterization of all microbial organisms and microbial communities.

WHY PARTICIPATE IN THE GMI PROFICIENCY TEST?

The proficiency test (PT) represents an important tool for the production of reliable laboratory results of consistently good quality within the area of DNA preparation, sequencing, and analysis (e.g. phylogeny).

WHAT IS OFFERED IN THE GMI PROFICIENCY TEST?

This inter-laboratory test performance is provided to ensure harmonization and standardization in whole genome sequencing and data analysis, with the aim to produce comparable data for the GMI initiative.

The GMI working group 4 (WG4) steered by the United States Food and Drug Administration, Microbiologics, and Technical University of Denmark has prepared this proficiency test. The PT consists of three parts, each of which are optional, and include assessing (1a) the laboratory's DNA preparation and sequencing procedures, (1b) the laboratory's sequencing output, and (2) the laboratory's procedure to cluster and distinguish samples based on the analysis of whole-genome-sequence data.

The proficiency test focuses on *Salmonella* Typhimurium, *Escherichia coli* and *Staphylococcus aureus*, and allows for sign-up for each species separately. Note that item 1a and item 1b are parallel; i.e. when signing up for 1a for one species, the participation in 1b is connected.

The three items consist of

> 1a) DNA extraction, purification, library-preparation, and whole-genome-sequencing of six bacterial cultures; two *Salmonella* Typhimurium strains, two *Escherichia coli* strains* and two *Staphylococcus aureus* strains. Participants will be requested to upload reads to an ftp-site and **optionally** also identify the Multi Locus Sequence Type (MLST) of the strains as well as the resistance genes present in the strains if that is something that is routinely done within the laboratory.

1b) Perform whole-genome-sequencing of pre-prepared DNA delivered by Working Group 4 of the GMI of the same six bacterial strains mentioned in clause 1a.

2) Phylogenetic/clustering analysis of three datasets each including fastq data from circa 20 genomes of *S.* Typhimurium, *E. coli* and *S. aureus*, respectively.

WHO SHOULD PARTICIPATE IN THE GMI PROFICIENCY TEST?

All laboratories of the GMI community performing whole-genome-sequencing and/or phylogenetic analysis are invited to participate, in particular, laboratories frequently submitting data to NCBI, EBI and DDBJ. For the GMI proficiency test pilot, however, participation is by invitation only.

COST FOR PARTICIPATING IN THE GMI PROFICIENCY TEST

There is no participation fee in the GMI proficiency test. Laboratories should, however, cover the expenses for parcel shipment if possible. If FedEx has 'Dangerous Goods-service' in your country or if you have a DHL-account number, please provide your FedEx or DHL import account number (for import of UN3373 Biological Substance Category B) in the sign-up form or, alternatively, to the PT Coordinator (please find contact information below). We need this information at this stage to save time and resources. Participating laboratories are responsible for all costs related to taxes or custom fees applied by their country as well as those related to the analysis.

HOW TO SIGN UP FOR THE GMI PROFICIENCY TEST

This link will open a sign-up webpage: http://www.globalmicrobialidentifier.org/Workgroups/GMI-Proficiency-Test-pilot-for-2014

In this webpage, you will be asked to provide the following information:

- Name of institute/organization and main contact person
- Complete mailing address for shipment of bacterial isolates and DNA
- Telephone and fax number, e-mail address
- FedEx or DHL import account number (if available)
- Items of the GMI PT you plan to participate in (item 1a, 1b and/or 2; *Salmonella*, *E. coli*, *S. aureus*)

If you experience any problem in the sign-up webpage please contact the GMI PT Coordinator Susanne Karlsmose: E-mail suska@food.dtu.dk; fax +45 3588 6341.

TIMELINE FOR SHIPMENT OF ISOLATES, DNA AND DATA-FILES

The bacterial isolates and the DNA will be shipped from DTU Food late June 2014 (pilot PT).

In order to minimize delays, **please send a valid import permit to the PT coordinator**. Please apply for a permit to receive the following bacterial cultures or DNA (according to your level of

participation): "UN3373, Biological Substance Category B": two *Salmonella* strains, one *Escherichia coli* strain*, two *S. aureus* strains.
Note: None of the bacterial cultures are enterotoxin producing.

AVAILABILITY OF PROTOCOLS

When isolates and DNA have been dispatched and data files been made available for download, protocols and all relevant information will be available for download from the website: http://www.globalmicrobialidentifier.org/Workgroups/About-the-GMI-Proficiency-Test-Pilot-2014.

DEADLINE FOR SUBMITTING RESULTS

Results must be submitted to an ftp-site as described in the protocol by **15th August 2014**. Full anonymity is ensured, and only the PT-organizers will have access to your results. An overall report summarizing the results will be published.

Deadline for sign-up for the GMI pilot PT 2014 is June 23rd 2014

* Due to issues in the preparations for the pilot PT, one set of *Escherichia coli* culture and DNA, only, will be included in the pilot PT, i.e. in total, two *Salmonella* Typhimurium, one *Escherichia coli* and two *Staphylococcus aureus*.

PROTOCOL for GMI proficiency test pilot, 2014

1 OVERVIEW AND OBJECTIVES 1

2 INTRODUCTION 2

3 OUTLINE OF THE GMI PT 3

3.1 Shipping, receipt and storage of bacterial strains 3

3.2 Using FTP to transfer files 3

3.3 Supplied test material 3

3.4 Procedure and analysis of test material 4

4 REPORTING OF RESULTS AND EVALUATION 5

1 OVERVIEW AND OBJECTIVES

The proficiency test pilot, 2014, consists of three general parts:

1a. DNA extraction, purification, library-preparation, and whole-genome-sequencing from **live cultures**

1b. Whole-genome-sequencing of **pre-prepared DNA**

2. Phylogenetic/clustering analysis of three **fastq datasets**

The main **objective** of this proficiency test is to quantify differences among laboratories in order to facilitate the development of reliable laboratory results of consistently good quality within the area of DNA preparation, sequencing, and analysis (e.g. phylogeny). This ensures that the discrepancies and differences among laboratories are known and will contribute to the standardization of whole genome sequencing and data analysis, with the aim to produce comparable data for the GMI initiative. A further objective is to assess and improve the uploaded data to databases such as NCBI, EBI and DDBJ.

2 INTRODUCTION

GMI is a global, visionary taskforce of scientists and other stakeholders who shares an aim of making novel genomic technologies and informatic tools available for improved global patient diagnostics, surveillance and research. This, by developing needs- and end-user-based data exchange and analysis tools for characterization of all microbial organisms and microbial communities.

The GMI working group 4 (WG4) steered by the United States Food and Drug Administration, Microbiologics, and Technical University of Denmark has prepared this proficiency test (PT). The PT consists of three parts, each of which are optional, and include assessing (1a) the laboratory's DNA preparation and sequencing procedures, (1b) the laboratory's sequencing output, and (2) the laboratory's procedure to analyse a whole-genome-sequencing dataset and distinguish between clonally related and sporadic genomes.

The proficiency test focuses on *Salmonella enterica* serovar Typhimurium, *Escherichia coli* strain and *Staphylococcus aureus*, and allows for sign-up for each species separately. Note that item 1a and item 1b are parallel; i.e. when signing up for 1a for one species, the participation in 1b is connected.

The three items consist of

1a) DNA extraction, purification, library-preparation, and whole-genome-sequencing of six bacterial cultures; two *Salmonella* Typhimurium strains, two *Escherichia coli* strain and two *Staphylococcus aureus* strains. Participants will be requested to upload reads to an ftp-site and **optionally** also identify the Multi Locus Sequence Type (MLST) of the strains as well as the resistance genes present in the strains if that is something that is routinely done within the laboratory.

1b) Whole-genome-sequencing of pre-prepared DNA of the same six bacterial strains mentioned in clause 1a.

2) Phylogenetic/clustering analysis of three datasets each including fastq data from circa 20 genomes of *S.* Typhimurium, *E. coli* and *S. aureus*, respectively.

Institutes/organizations which signed up to participate will receive the PT-material (bacterial strains, DNA and/or the login for download of datasets) according to the information reported in the sign-up form.

To achieve the objective of assessing and improving the uploaded data to databases such as NCBI, EBI and DDBJ, the laboratory work analysis performed for this PT should be done by using the methods routinely used in your laboratory.

3 OUTLINE OF THE GMI PT

3.1 Shipping, receipt and storage of bacterial strains

In June 2014, around 10 laboratories located worldwide will receive a parcel containing the two *S.* Typhimurium strains, two *E. coli* strains and two *S. aureus* strains together with corresponding purified DNA (according to information reported in the sign-up form). All bacterial strains and DNA are shipped as UN3373, Biological substance category B. Those who signed up for item 2 (phylogenetic analysis) will receive information and login for downloading the three datasets.

Please confirm receipt of the parcel through the confirmation form enclosed in the shipment.

The bacterial strains are shipped lyophilised as KwikStik's (see below for additional info on handling). On arrival, the KwikStik's must be refrigerated until handling in the laboratory.

The bacterial DNA is shipped as dried samples using a DNA stabilizing agent (DNAstable® *Plus*, Biomatrica). On arrival, either rehydrate your sample and store the liquid samples at room temperature in closed tubes, to prevent evaporation. Or store the dried samples in either

(a) a dry storage cabinet at room temperature (15-25°C or 59-77°F) or
(b) a heat-sealed, moisture-barrier bag along with a silica gel desiccant pack.

3.2 Using FTP to transfer files

For download of fastq files for item 2 and for upload of results, an ftp-server is used. The proficiency test organizer will provide each participant with username and login for this purpose. The ftp-site which will be used for this purpose is cgebase.cbs.dtu.dk. For information on how to transfer files, please see Appendix 1.

3.3 Supplied test material

3.3.1 Item 1a; Bacterial cultures

The procedure for reconstitution of the bacterial cultures should follow the manufacturer's procedures as presented in the instructional video or the written instructions on their website (see http://microbiologics.com/s.nl/sc.7/category.98564/.f or http://microbiologics.com/Support-Center/KWIK-STIK-trade).

MSDS for Kwik stik are found here: http://microbiologics.com/Support-Center/Lyophilized-Microorganism-Preparations.

The bacterial cultures supplied have been sequenced multiple times and the genomes have been closed. Therefore, the PT-organizers encourage participants to maintain these bacterial strains in their strain collection and apply them as part of future internal quality control.

3.3.2 Item 1b; DNA

The supplied DNA has been stabilized by DNA Stable®*plus* (http://www.biomatrica.com/media/dnastable%20Plus/3004-0112.pdf). Each vial contains a minimum of 3ug DNA. Before use, the samples should be re-suspended in 100 µl water or aqueous buffer and mixed by gentle pipetting or vortexing for 15 min (according to above mentioned protocol). Rehydrated samples can be stored at room temperature and used directly in downstream application.

3.3.3 Item 2; Fastq data set

Three datasets, one for each of *S.* Typhimurium*, E. coli* and *S. aureus*, will be available for download from the ftp-site 'cgebase.cbs.dtu.dk'. Login to the ftp-site will be provided directly to each participant. Each dataset will consist of the original fastq files (i.e., whole genome sequence data) from circa 20 samples for phylogenetic cluster analysis based on a tool of the laboratory's own choice; SNP-calling, gene-by-gene, etc.

3.4 Procedure and analysis of test material

3.4.1 Item 1a and 1b; Bacterial cultures and DNA

Subculture the bacterial strains on a relevant growth medium of the laboratory's own choice and incubate. Following incubation and assessment of purity of the bacterial cultures, perform DNA extraction and whole-genome-sequencing according to the laboratory's standard procedure.

For the purified PT-DNA received, perform whole-genome-sequencing according to the laboratory's standard procedure.

For both bacterial cultures and DNA (items 1a and 1b), register relevant information related to the methods applied via https://www.surveymonkey.com/s/pilot-PT_2014_bacterial_cultures_and_DNA (also see Appendix 2). Upload the non-assembled sequence data obtained (e.g., fastq-files, sff files) for each of the bacterial cultures and DNA-samples to the ftp-site (cgebase.cbs.dtu.dk) using your individual login and password. Appendix 2 also describes the submission of results obtained when analyzing the sequences as regards the detected antimicrobial resistance genes and as regards the Multi Locus Sequence Type of the bacterial strain.

For both bacterial cultures and DNA (items 1a and 1b), the submitted sequence data (fastq-files) will be evaluated according to the following specific quality markers: Read length (bp), N50 (bp), total number of contigs and total length of sequence (bp) including percentage of reference genome covered. In addition, the PT-organizers will assemble the submitted reads and compare

these assemblies 1) towards the relevant closed genome to assess the sequence error rate and coverage of the scaffold and 2) between the obtained sequences in items 1a and 1b.

3.4.2 Item 2; Fastq data set

The three fastq datasets should be downloaded from the ftp-site. They are organized into three different .zip archives appropriately labeled with the taxon they represent. Within each archive the participant will find the paired-end reads, a tab-delimited file containing info about the isolates, and a pdf document with questions to be answered based on the results of the analysis. The objective associated with this dataset is to assess the variability of laboratories in the clusters identified through the analysis of next-generation sequencing data. As such, the participant should employ their preferred method for constructing a matrix (e.g., gene, SNP, presence/absence, etc.) and for clustering samples (e.g., distance-, maximum-likelihood-, Bayesian-based).

Assessment of the submitted results from the analysis of the datasets is based on two criteria: 1) the concordance among laboratories in their answers to the questions within the pdf document and 2) the concordance between participants' in the information content contained in the matrix and the relationships among samples from the clustering analyses (i.e., the topology).

Specifically, four types of files should be submitted:

For each dataset:

1. The DNA sequence matrix used for clustering (e.g., a fasta, phylip, or nexus formatted file)
2. If relevant, also submit the distance matrix file
3. The clusters themselves (e.g., a newick or nexus formatted tree file)
4. A spreadsheet listing the obtained coverage and the number of bases sequenced (download template from the 'shared' ftp-site)

For each strain:

5. For each strain, one file containing the assembled contigs should be submitted. The filename should consist of username, strain name and type of file, all in lowercase letters, e.g. 'username_strainname_typeoffile.xxx'

From the files containing assembled contigs, the proficiency test organizers will calculate N50, number of contigs, size of the chromosome and the longest contig.

Via the Internet-based survey (https://www.surveymonkey.com/s/pilot-PT_2014_FASTQ_dataset; see also Appendix 3), answers should be submitted to the questions related to the Fastq data set section.

4 REPORTING OF RESULTS AND EVALUATION

The results, which should be captured and entered into the Internet-based survey (https://www.surveymonkey.com/s/pilot-PT_2014_bacterial_cultures_and_DNA and https://www.surveymonkey.com/s/pilot-PT_2014_FASTQ_dataset), are listed in Appendix 2 and 3.

Results must be submitted electronically no later than 15 August 2014. Immediately after this date, the survey will be closed and results submitted to the Internet-based survey and to the ftp-site will be evaluated. Delayed submission of results will not be accepted. All submitted results will be summarized in a report which will be publicly available. Individual results will be anonymized. For the evaluation of the results, no official GMI quality threshold is currently available and therefore no acceptance limit has been defined for this proficiency test.

We are looking forward to receiving your results.

If you have any questions or concerns, please do not hesitate to contact us:

In relation issues related to dry-lab issues, please contact:

James Pettengill
FDA
Maryland, US
E-mail: James.Pettengill@fda.hhs.gov

In relation to other issues, e.g. organizational issues, please contact the EQAS Coordinator:

Susanne Karlsmose
National Food Institute, Technical University of Denmark
Kemitorvet, Building 204 ground floor, DK-2800 Kgs. Lyngby - DENMARK
Tel: +45 3588 6601, Fax: +45 3588 6341
E-mail: suska@food.dtu.dk

___ ___ ___

Appendix 1

Using FTP to transfer files

FTP is an acronym for File Transfer Protocol and is used to transfer files between computers on a network. To access the folder for upload or download of files, do as described below.

Obtain access to upload or download files by using the relevant login provided by the proficiency test organizer.

Using a Windows-computer:

Open the Documents folder, and type 'ftp://cgebase.cbs.dtu.dk/' in the Address bar.
Enter you username and password, click "Log on".

Using a Mac-computer:

FileZilla FTP client:
- Download and install FileZilla (https://filezilla-project.org/)
- Host:cgebase.cbs.dtu.dk
- Type username and password
- Connect

Or

Finder Mac application:
- In the Finder, choose Go > "Connect to Server," and wait for the pop-up window to show up.
- Specify server address ftp://cgebase.cbs.dtu.dk and click "Connect"
- In the new pop-up window enter you username and password, click "Connect"

GMI Proficiency Testing Pilot (PT) 2014 - bacterial cultures and DNA

Introduction

This survey seeks to capture info on participants' sequence procedures and specifications in relation to the tested bacterial cultures and DNA.

The survey consists of five sections, collecting information on
1. User Information and Sample Storage
2. Bacterial Culture; DNA Isolation, Handling and Processing
3. Received DNA; Handling and Processing
4. Sequencing
5. Analysis of sequences; MLST and antimicrobial resistance genes

If you have any questions or feedback for the submission of data via this survey, please contact the PT Coordinator, Susanne Karlsmose (suska@food.dtu.dk), at the Technical University of Denmark.

Note: An asterisk (*) indicates a question that requires an answer.

GMI is a global, visionary taskforce of scientists and other stakeholders who shares an aim of making novel genomic technologies and informatics tools available for improved global patient diagnostics, surveillance and research, by developing needs- and end-user-based data exchange and analysis tools for characterization of all microbial organisms and microbial communities.

*1. Institute name / Organization name

*2. Department name

*3. Name of person responsible for the handling of the PT-material

*4. Dates in relation to the handling of the PT-material (date for upload of sequence data)

	DD / MM / YYYY
PT-material reception date	/ /
Test start of processing the bacterial cultures	/ /
Completion of processing the bacterial cultures	/ /
Test start of processing the DNA	/ /
Completion of processing the DNA (upload of sequence data)	/ /

GMI Proficiency Testing Pilot (PT) 2014 - bacterial cultures and DNA

***5. Storage conditions of the bacterial cultures in the time between reception and processing:**
(please select one answer)

- ◯ -80˚C
- ◯ -20˚C
- ◯ 4˚C
- ◯ Room temperature
- ◯ No storage time
- ◯ We did not receive bacterial cultures for this PT
- ◯ Other

If other, please define

***6. Storage conditions of the DNA in the time between reception and processing:**
(please select one answer)

- ◯ -80˚C
- ◯ -20˚C
- ◯ 4˚C
- ◯ Room temperature
- ◯ No storage time
- ◯ Other

If other, please define

BACTERIAL CULTURES received

***7. How were the bacterial cultures cultivated:**

7.1 - Type of agar media/liquid broth:

7.2 - Incubation time:

7.3 - Incubation temperature:

GMI Proficiency Testing Pilot (PT) 2014 - bacterial cultures and DNA

*8. For the Gram-negative bacterial cultures; DNA extraction procedure:

8.1 - If manual extraction; kit used:

8.2 - If manual extraction, modifications to kit protocol:

8.3 - If automatic extraction; robot used:

8.4 - If automatic extraction; specific protocol:

8.5 - If automatic extraction; modifications to protocol:

*9. For the Gram-positive bacterial cultures; DNA extraction procedure:

9.1 - If manual extraction; kit used:

9.2 - If manual extraction, modifications to kit protocol:

9.3 - If automatic extraction; robot used:

9.4 - If automatic extraction; specific protocol:

9.5 - If automatic extraction; modifications to protocol:

10. For bacterial cultures, DNA concentration (ng/µl) prior to library preparation was measured on (please select one answer)

- ◯ Qubit
- ◯ Nanodrop
- ◯ DNA concentration not measured
- ◯ Other

If other, please specify:

*11. Measure of DNA concentration (ng/µl)

11.1 GMI14-001-BACT (Salmonella)

11.2 GMI14-002-BACT (Salmonella)

11.3 GMI14-003-BACT (E. coli)

11.4 GMI14-004-BACT (E. coli)

11.5 GMI14-005-BACT (S. aureus)

11.6 GMI14-006-BACT (S. aureus)

GMI Proficiency Testing Pilot (PT) 2014 - bacterial cultures and DNA

12. Total DNA amount (microgram)

12.1 GMI14-001-BACT (Salmonella)

12.2 GMI14-002-BACT (Salmonella)

12.3 GMI14-003-BACT (E. coli)

12.4 GMI14-004-BACT (E. coli)

12.5 GMI14-005-BACT (S. aureus)

12.6 GMI14-006-BACT (S. aureus)

13. For bacterial cultures, DNA quality (e.g. RIN, 260/280 ratio and/or 260/230 ratio) prior to library preparation was measured on (please select one answer)

- ◯ Bioanalyser
- ◯ Nanodrop
- ◯ DNA quality not measured
- ◯ Other

If other, please specify:

14. Measure of DNA quality (e.g. RIN or 260/280 ratio)

14.1 GMI14-001-BACT (Salmonella)

14.2 GMI14-002-BACT (Salmonella)

14.3 GMI14-003-BACT (E. coli)

14.4 GMI14-004-BACT (E. coli)

14.5 GMI14-005-BACT (S. aureus)

14.6 GMI14-006-BACT (S. aureus)

15. If relevant; measure of DNA quality (260/230 ratio):

15.1 GMI14-001-BACT (Salmonella)

15.2 GMI14-002-BACT (Salmonella)

15.3 GMI14-003-BACT (E. coli)

15.4 GMI14-004-BACT (E. coli)

15.5 GMI14-005-BACT (S. aureus)

15.6 GMI14-006-BACT (S. aureus)

DNA received

GMI Proficiency Testing Pilot (PT) 2014 - bacterial cultures and DNA

16. For the DNA received, DNA concentration (ng/µl) prior to library preparation was measured on (please select one answer)

- ◯ Qubit
- ◯ Nanodrop
- ◯ DNA concentration not measured
- ◯ Other

If other, please specify:

17. Measure of DNA concentration (ng/µl)

17.1 GMI14-001-DNA (Salmonella)	
17.2 GMI14-002-DNA (Salmonella)	
17.3 GMI14-003-DNA (E. coli)	
17.4 GMI14-004-DNA (E. coli)	
17.5 GMI14-005-DNA (S. aureus)	
17.6 GMI14-006-DNA (S. aureus)	

18. Total DNA amount (microgram)

18.1 GMI14-001-DNA (Salmonella)	
18.2 GMI14-002-DNA (Salmonella)	
18.3 GMI14-003-DNA (E. coli)	
18.4 GMI14-004-DNA (E. coli)	
18.5 GMI14-005-DNA (S. aureus)	
18.6 GMI14-006-DNA (S. aureus)	

19. For the DNA received, DNA quality (e.g. RIN, 260/280 ratio and/or 260/230 ratio) prior to library preparation was measured on (please select one answer)

- ◯ Bioanalyser
- ◯ Nanodrop
- ◯ DNA quality not measured
- ◯ Other

If other, please specify:

GMI Proficiency Testing Pilot (PT) 2014 - bacterial cultures and DNA

20. Measure of DNA quality (e.g. RIN or 260/280 ratio)

20.1 GMI14-001-DNA (Salmonella)

20.2 GMI14-002-DNA (Salmonella)

20.3 GMI14-003-DNA (E. coli)

20.4 GMI14-004-DNA (E. coli)

20.5 GMI14-005-DNA (S. aureus)

20.6 GMI14-006-DNA (S. aureus)

21. If relevant; measure of DNA quality (260/230 ratio):

21.1 GMI14-001-DNA (Salmonella)

21.2 GMI14-002-DNA (Salmonella)

21.3 GMI14-003-DNA (E. coli)

21.4 GMI14-004-DNA (E. coli)

21.5 GMI14-005-DNA (S. aureus)

21.6 GMI14-006-DNA (S. aureus)

SEQUENCING

22. What protocol was used to prepare the sample library for sequencing? For commercial kits please provide the full kit name, item number, and lot number if possible. For noncommercial kits please provide a citation for the protocol, or submit a summary of the protocol. Please note any deviations from the kit or cited protocol

For commercial kits; full kit name:

For commercial kits; item number:

For commercial kits; lot number:

For noncommercial kits; citation for the protocol:

For noncommercial kits; summary of the protocol:

Deviations from the kit or cited protocol

GMI Proficiency Testing Pilot (PT) 2014 - bacterial cultures and DNA

***23. Please indicate the sequencing platform you used in the proficiency test (please select one answer)**

- ◯ Ion Torrent PGM
- ◯ Ion Torrent Proton
- ◯ Genome Sequencer Junior System (454)
- ◯ Genome Sequencer FLX System (454)
- ◯ Genome Sequencer FLX+ System (454)
- ◯ PacBio RS
- ◯ PacBio RS II
- ◯ HiScanSQ
- ◯ HiSeq 1000
- ◯ HISeq 1500
- ◯ HiSeq 2000
- ◯ HiSeq 2500
- ◯ Genome Analyzer Iix
- ◯ MiSeq Benchtop Sequencer
- ◯ ABI SOLiD
- ◯ other

If other, please specify

24. Sequencing details #1 (please select one answer)

- ◯ Single-end
- ◯ Paired-end
- ◯ Not relevant

25. Sequencing details #2:

For the sequencing, the read length (bp) was set to be (expected read length)

***26. Reads trimmed before upload (please select one answer)**

- ◯ Yes
- ◯ No

If trimmed, which tool was applied (in the text box below, please insert name and URL (if possible))

GMI Proficiency Testing Pilot (PT) 2014 - bacterial cultures and DNA

27. For this item (1a and 1b) in the proficiency test, assembly is not requried. If, however, you were to assemble your sequences, which assembly tool would you apply? in the text box below, please insert name and URL (e.g. Velvet, https://www.ebi.ac.uk/~zerbino/velvet/, open access)

Assembly tool:

ANALYSIS of sequences

28. If any, which method was used to characterize or differentiate isolates?

- [ ] MLST
- [ ] Allele-based
- [ ] Gene-by-gene-based
- [ ] SNP-based
- [ ] None

Other (please specify)

29. Did you determine which antimicrobial resistance genes were present in the sequenced DNA?

- ◯ Yes
- ◯ No

30. For the DNA from the received bacterial culture, if analysis was performed based on the sequence analysis, which MLST-type does the isolate belong to, or which alleles characterize the isolate?

30.1 GMI14-001-BACT (Salmonella)

30.2 GMI14-002-BACT (Salmonella)

30.3 GMI14-003-BACT (E. coli)

30.4 GMI14-004-BACT (E. coli)

30.5 GMI14-005-BACT (S. aureus)

30.6 GMI14-006-BACT (S. aureus)

GMI Proficiency Testing Pilot (PT) 2014 - bacterial cultures and DNA

31. For the DNA from the received bacterial culture, if analysis for antimicrobial resistance genes was performed based on the sequence analysis, which antimicrobial resistance genes does the isolate harbour?

31.1 GMI14-001-BACT (Salmonella)

31.2 GMI14-002-BACT (Salmonella)

31.3 GMI14-003-BACT (E. coli)

31.4 GMI14-004-BACT (E. coli)

31.5 GMI14-005-BACT (S. aureus)

31.6 GMI14-006-BACT (S. aureus)

32. For the received DNA, if MLST-analysis was performed based on the sequence analysis, which MLST-type does the isolate belong to, or which alleles characterize the isolate?

32.1 GMI14-001-DNA (Salmonella)

32.2 GMI14-002-DNA (Salmonella)

32.3 GMI14-003-DNA (E. coli)

32.4 GMI14-004-DNA (E. coli)

32.5 GMI14-005-DNA (S. aureus)

32.6 GMI14-006-DNA (S. aureus)

33. For the received DNA, if analysis for antimicrobial resistance genes was performed based on the sequence analysis, which antimicrobial resistance genes does the isolate harbour?

33.1 GMI14-001-DNA (Salmonella)

33.2 GMI14-002-DNA (Salmonella)

33.3 GMI14-003-DNA (E. coli)

33.4 GMI14-004-DNA (E. coli)

33.5 GMI14-005-DNA (S. aureus)

33.6 GMI14-006-DNA (S. aureus)

34. For the detection of the Multi Locus Sequence Type, which tool did you apply? in the text box below, please insert name and URL (e.g. MLST 1.7 (MultiLocus Sequence Typing), http://cge.cbs.dtu.dk/services/MLST/, open access)

Tool for detection of MLST:

35. For the detection of the resistance genes harboured in the seqences, which tool did you apply? in the text box below, please insert name and URL (e.g. ResFinder, http://cge.cbs.dtu.dk/services/ResFinder/, open access)

Tool for detection of resistance genes:

GMI Proficiency Testing (PT) Pilot 2014 - FASTQ dataset

Introduction

With this survey we seek to capture info in relation to the fastq data set component of the GMI pilot PT.

If you have any questions or feedback for the submission of data via this survey, please contact the PT Coordinator, Susanne Karlsmose (suska@food.dtu.dk), at the Technical University of Denmark.

Note: An asterisk (*) indicates a question that requires an answer.

GMI is a global, visionary taskforce of scientists and other stakeholders who shares an aim of making novel genomic technologies and informatics tools available for improved global patient diagnostics, surveillance and research, by developing needs- and end-user-based data exchange and analysis tools for characterization of all microbial organisms and microbial communities.

***1. Institute name / Organization name**

***2. Department name**

***3. Name(s) of person(s) responsible for the analysis**

FASTQ data set

4. Were reads quality filtered before conducting the analysis?
(please select one answer)

- ◯ Yes
- ◯ No

5. If reads were quality filtered, please provide the name of the program

6. For the assembly of the contigs, which tool did you apply?
In the text box below, please insert name and URL (e.g. Velvet, https://www.ebi.ac.uk/~zerbino/velvet/)

***7. What kind of methodology for phylogeny construction did you apply?**

- ◯ SNPs
- ◯ Methodology other than SNPs

If methodology other than SNPs (please specify):

GMI Proficiency Testing (PT) Pilot 2014 - FASTQ dataset

If applying SNPs, go to question 8,
If not applying SNPs, go to question 11

8. Which reference genome did you use for SNP calling? If you used a reference-free approach, please indicate 'none'.

7.1 - S. Typhimurium	
7.2 - E. coli	
7.3 - S. aureus	

9. Which quality criteria did you use for SNP calling? (e.g. % of mapped reads and minimum coverage to define variant).

8.1 - S. Typhimurium	
8.2 - E. coli	
8.3 - S. aureus	

10. Which criteria did you use for SNPs filtering:

9.1 - Filter SNPs with excess coverage (i.e. repetitive regions):	
9.2 - Did you filter SNPs occurring in a cluster (a.k.a. pruning) (indicate 'yes' or 'no'):	
9.3 - Which definition of the cluster did you use (i.e. ≥3 SNPs in 1000 base pairs (bp):	
9.4 - Other, please specify:	

11. Which program did you use to build your tree (e.g., MEGA, MrBayes, PAUP*, GARLI, RAxML, etc)?

12. Which algorithm did you use to build your tree (e.g., Neighbor-joining, UPGMA, Bayesian, maximum-likelihood, etc)?

Please upload to the ftp-site your DNA sequence matrix as a fasta alignment file, and, if relevant, the distance matrix file (as also described in the protocol)

13. Do you calculate the number of contigs (please select one answer)

- ◯ Yes
- ◯ No

14. Do you filter out contigs below a certain size (please select one answer)

- ◯ Yes
- ◯ No

If yes, indicate minimum size

GMI Proficiency Testing (PT) Pilot 2014 - FASTQ dataset

15. Do you calculate N50 (please select one answer)

- ◯ Yes
- ◯ No

16. Do you calculate N50 before or after contig size filtering? (please select one answer)

- ◯ Before
- ◯ After
- ◯ No filtering

17. Do you calculate the size of the chromosome (please select one answer)

- ◯ Yes
- ◯ No

18. Do you calculate coverage (please select one answer)

- ◯ Yes
- ◯ No

19. If you calculate coverage, describe how you estimated genome size

As also described in the protocol, please upload to the ftp-site a file containing a table listing the coverage found for each of the 20, 21, 24 sequences of S. Typhimurium, E. coli and S. aureus. List the code of the strain followed by the obtained coverage.
Also, please upload a file containing a table listing the number of bases you sequenced for each of the 20, 21, 24 sequences of S. Typhimurium, E. coli and S. aureus. List the code of the strain followed the number of bases you sequenced.

20. For verification of species, which tool did you apply? In the text box below, please insert name and URL (e.g. KmerFinder 1.2, http://cge.cbs.dtu.dk/services/KmerFinder)

21. Could you verify the species?

- ◯ We did not attempt to verify species
- ◯ Yes, for all S. Typhimurium
- ◯ Yes, for all E. coli
- ◯ Yes, for all S. aureus
- ◯ No, for some S. Typhimurium
- ◯ No, for some E. coli
- ◯ No, for some S. aureus

If no, please indicate why

GMI Proficiency Testing (PT) Pilot 2014 - FASTQ dataset

22. Can you call a Multi Locus Seqeunce Type (MLST) (please select one answer)

- ◯ Yes
- ◯ No
- ◯ We are not interested in MLST

23. Please describe/interpretation the tree that you obtained based on the following questions; 1) How does the analysis indicate that the isolates are linked? 2) How many clusters could be identified? 3) How could the clusters be defined?

S. Typhimurium	
E. coli	
S. aureus	

GMI proficiency test pilot, 2014

Id: [Id]
[Name]
[Name of institute/organization]
[Country]

Kgs. Lyngby, Denmark, June 2014

Dear [Name],

Please find enclosed the bacterial cultures and DNA for the GMI proficiency test pilot, 2014. The bacterial strains are shipped lyophilised as KwikStik's (see further information in the protocol). On arrival, they must be refrigerated until handling in the laboratory. The bacterial DNA is shipped as dried samples using a DNA stabilizing agent (DNAstable® *Plus*, Biomatrica). On arrival, either rehydrate your sample and store the liquid samples at room temperature in closed tubes, to prevent evaporation. Or store the dried samples in either (a) a dry storage cabinet at room temperature (15-25°C or 59-77°F) or (b) a heat-sealed, moisture-barrier bag along with a silica gel desiccant pack.

The following documents and information relevant for the GMI proficiency test pilot are available on the GMI website (see http://www.globalmicrobialidentifier.org/Workgroups/About-the-GMI-Proficiency-Test-Pilot-2014):

- Protocol for GMI pilot proficiency test 2014
- Link to SurveyMonkey – submission of results for bacterial cultures and DNA
- Link to SurveyMonkey – submission of results for FASTQ data sets

In the protocol, you will find detailed description for testing the bacterial cultures (item 1a), the corresponding DNA (item 1b), and the analysis of the provided datasets. Additionally, you will find instructions for submission of results and sequences. To access the ftp-server to/from which sequence information can be up- and downloaded, you need the username and password listed below.

Results must be submitted electronically no later than 15 August 2014

Please acknowledge receipt of this parcel immediately upon arrival (see enclosed 'Confirmation Form').

Do not hesitate to contact us for further information,

Susanne Karlsmose
EQAS-Coordinator

For upload:
Your username:[user_upload]
Your password: [password_upload]

For download:
Your username: [user_download]
Your password: [password_download]

Please keep this document
Your usernames and passwords will not appear in other documents

Descriptions of how clusters were defined as part of the Dry lab component

S. Typhimurium
1) S2 and S12 are closely related to each other. The remaining strains are closely related to each other, except S8 (which is further away from S2&S12 than from the other strains but seems to still be only distantly related to the other strains) 2) 2 clusters. 3) Cluster identification via intra- and intercluster distances using CTree (http://www.phylogenetictrees.com/ctree.php)
1) There is a breadth of diversity among isolates; 2) there is one large cluster with possible two other isolates forming a smaller cluster; 3) clusters were defined on relative SNP differences among isolates
Three clusters of closely related isolates, one with 2 isolates, one with 4 isolates that included the reference (LT2)
3 clusters, 2 closely related clusters, then 1 distantly related cluster
Isolates from the fastq dataset formed three clusters. One cluster contained 15 isolates, 14 of which were extremely closely related, indicating they could be part of an outbreak. The second four-isolate cluster contained 2 subsets each containing a pair of closely related isolates. Isolate FSW0035_S11 did not cluster with any of the other isolates.
1) Well supported clades with small snp distance. 2) 3 clusters. 3) Well supported distinct clades with 0 to 39 interstrain snp distance
1) based on bootstrap 2) 6 clusters 3) based on phylogenetic analysis, MLST and resistance genes

E. coli
1) S1 and S10 are closely related to each other. The remaining strains are closely related to each other. 2) 2 clusters. 3) Cluster identification via intra- and intercluster distances using CTree (http://www.phylogenetictrees.com/ctree.php)
1) there is a breadth of diversity among isolates; 2) in terms of foodborne outbreak analysis, there only appear to be two samples that would be considered linked; 3) clusters were defined on relative SNP differences among isolates
two clusters one with two isolates, and the rest belonging to a single cluster with 40 or fewer SNPs differentiating the strains
2 distantly related isolates and 1 main cluster of closely related isolates
Because we included the isolate 004_E. coli that we sequenced in part one, the resolution of the tree was low, so although we could detect three distinct clusters we could not infer the relationships of strains within the clusters
1) Well supported clades with small snp distance. 2) 4 clusters. 3) Well supported distinct clades with 1 to 63 interstrain snp distance
1) based on bootstrap 2) 5 clusters 3) based on phylogenetic analysis, MLST and resistance genes

S. aureus 1) Closely related to each other: (M_S5,M_S9), (H_S8,M_S3,H_S5,H_S9,H_S10,H_S7,H_S4,H_S6,H_S12,H_S2,H_S1), (M_S12,M_S7,M_S4,M_S1,M_S18,M_S11), (M_S3,M_S6),(M_S11),(H_S15),(H_S11). 2) 6 clusters. 3) Cluster identification via intra- and intercluster distances using CTree (http://www.phylogenetictrees.com/ctree.php)
1) there is a breadth of diversity among isolates; 2) one large cluster likely exists perhaps with a second smaller one with about nine samples not belonging to any cluster; 3) clusters were defined on relative SNP differences among isolates
three clusters, one distantly related isolate that clustered with the reference, then one cluster of isolates with less then 3 SNPs differentiating the strains, and a second more diverse cluster with less then 610 SNP differentiating the isolates, this more diverse cluster included three identical isolates
2 main clusters with 1 group of closely related isolates in one of the clusters
Isolates were divided into two distinct groups, one containing eight isolates and the other containing 16. The latter group contained a cluster of 11 very closely related isolates that could belong to a single outbreak, the five other isolates in this group all distinct and likely to represent un-linked cases. The group containing eight isolates forms two distinct sub-clusters.
1) Well supported clades with small snp distance. 2) 3 clusters. 3) Well supported distinct clades with 0 to 6 interstrain snp distance
1) based on bootstrap 2) 4 clusters 3) based on phylogenetic analysis, MLST and resistance genes

National Food Institute
Technical University of Denmark
Kemitorvet, 204
DK - 2800 Kgs Lyngby

Tel. 35 88 70 00
Fax 35 88 70 01

www.food.dtu.dk

ISBN: 978-87-93109-78-0