REPORT ON THE 2nd PROFICIENCY TEST TRIAL FOR THE GLOBAL MICROBIAL IDENTIFIER (GMI) INITIATIVE, YEAR 2016

Photo: Colourbox

DTU Food
National Food Institute

ENSURING THE QUALITY OF BACTERIAL WHOLE GENOME SEQUENCING: REPORT ON THE 2nd PROFICIENCY TEST TRIAL FOR THE GLOBAL MICROBIAL IDENTIFIER (GMI) INITIATIVE, YEAR 2016

Rene S. Hendriksen[1], Oksana Lukjancenko[1], Susanne Karlsmose Pedersen[1], Jose Luis Bellod Cisneros[2], Lukasz Dariusz Dynowski[2], Ole Lund[2], Pimlapas Leekitcharoenphon[1], Rolf Sommer Kaas[1], Berith Elkær Knudsen[1], Inge Marianne Hansen[1], Jacob Dyring Jensen[1], Hanne Mordhorst[1], Vitali Sintchenko[3,4], William J. Wolfgang[5], Henrik Torkil Westh[6], Jacob Moran-Gilad[7], William Hsiao[8], Isabel Cuesta[9], Sara Monzón[9], Angel Zaballos[10], Anthony Underwood[11], Errol Strain[12], James Pettengill[12], Frank M. Aarestrup[1] 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]Technical University of Denmark, Bioinformatics, Department of Bio and Health Informatics, Kgs. Lyngby, Denmark

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

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

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

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

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

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

[9]National Center for Microbiology, Bioinformatics Unit, Institute of Health Carlos III Carretera Majadahonda, Madrid, Spain

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

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

[12]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 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 affiliated venues, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

1. edition, October 2017
Copyright: National Food Institute, Technical University of Denmark
Photo: Colourbox
ISBN: 978-87-93565-14-2

The report is available at
http://www.globalmicrobialidentifier.org/workgroups/gmi-proficiency-test-reports
National Food Institute
Technical University of Denmark
Kemitorvet
Building 204
DK-2800 Kgs. Lyngby
Denmark
Tel: +45 35 88 70 00

Contents

List of Abbreviations

AMR	**Antimicrobial resistance**
BWA	**Burrows-Wheeler Aligner**
CGE	**Center for Genomic Epidemiology**
C. jejuni	***Campylobacter jejuni***
C. coli	***Campylobacter coli***
DNA	**Deoxyribonucleic acid**
DTU	**Technical University of Denmark**
GMI	**Global Microbial Identifier**
HGAP	**Hierarchical Genome Assembly Process**
IATA	**International Air Transportation Association**
K. pneumoniae	***Klebsiella pneumoniae***
L. monocytogenes	***Listeria monocytogenes***
MDR	**Multidrug Resistant**
MLST	**Multi Locus Sequence Typing**
PT	**Proficiency test**
QC	**Quality Control**
SMRT	**Single-molecule real-time**
SNP	**Single Nucleotide Polymorphism**
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. Executive summary

The main objective of the proficiency test (PT) is to facilitate the production of reliable laboratory results of consistently good quality within the area of whole genome sequencing (WGS). The PT consists of two components, “wet-lab” and “dry-lab” targeting various bacterial pathogens. Only the wet-lab part will be addressed in this report. The wet-lab component assesses 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.

In 2016, two strains each of *Campylobacter* (GMI16-001, GMI16-002) *Listeria monocytogenes* (GMI16-003, GMI16-004), and *Klebsiella pneumoniae* (GMI16-005, GMI16-006) were selected for the wet-lab component. A total of 46 laboratories (two laboratories split the participation in between two departments) in 22 countries, four continents uploaded data for the PT.

A total of 13 (28%) out of 46 laboratories were identified to perform unsatisfactory and determined outliers for one or more of the QC parameters. This included the following laboratories: #71, #75, #79, #80, #83, #86, #91, #93, #104, #105, #107, #115, and #120. The poor performance was either a result of laboratory contamination of the culture and/or DNA or the sequencing itself which also appears to be somehow related to specific platforms. Tentative QC thresholds were determined for a number of parameters not depended on the availability of closed genomes of the test strains.

Most of the participants were able to identify the correct MLST with minor problems related to submission of the data mixing up the result of one genome with other genomes. The identification of the AMR genes provided a bit more information related to quality. Similarly, large discrepancies observed were a result of the same issue mixing up the data whereas minor differences related to the expected AMR profile were more related to sequence quality. The GMI PT will continue also in 2017 and might contain minor changes in the overall analysis.

2. Introduction

The main objective of the 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 target audience as well as the bacterial pathogens of interest. In addition, the survey captured the information about the current quality markers being employed to ensure high quality sequencing data (2). The results of this survey were utilized to create the foundation of the present PT.

Specifically, the PT aims at evaluating the consistency and robustness of Global Microbial Identifier (GMI) members' and others' ability to perform deoxyribonucleic acid (DNA) extraction, library preparation, WGS or a combination thereof, 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. To meet the objective, the laboratory work and analyses should be performed using the methods routinely employed in the individual laboratories.

The PT consists of two components, "wet-lab" and "dry-lab" targeting various bacterial pathogens. Only the wet-lab part will be addressed in this report. The wet-lab component assesses 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 main organizers of the GMI 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), and potential assisting members of the TAG. All summary conclusions are made public. The tentative goals 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).

3. Materials and Methods

3.1 Participating laboratories

A pre-notification to announce the 3rd GMI proficiency test was distributed on the 12 July 2016 by e-mail to the GMI members as well as posted on the GMI web site. A total of 50 members signed up for the wet-lab component and 47 laboratories participated by submitting sequences and metadata. Some of the laboratories, however, only took part in testing a subset of the target organisms after agreement with the PT organizers. The results from the laboratories participating in the wet-lab component are presented and evaluated in this report.

3.2 Strains

In 2016, two strains each of *Campylobacter* (GMI16-001, GMI16-002) *Listeria monocytogenes* (GMI16-003, GMI16-004), and *Klebsiella pneumoniae* (GMI16-005, GMI16-006) were selected for the wet-lab component. In a GMI end-user analysis of what species to target, *Campylobacter* and *Listeria* were indicated to be of interest (2). *Campylobacter*, a foodborne pathogen carrying a significant burden worldwide was selected for this PT due its heterogeneity with many repeats and rearrangements in the genomes and *Listeria*, the most lethal of foodborne pathogens and for being genetically homogenous with limited repeats and rearrangements. One of the *Listeria* strains belonged to known virulent MLST type, ST-2 (GMI16-003) as well as a less virulent ST-121 (GMI16-004). We also included *Klebsiella* due to its many AMR genes with the purpose of evaluating if the detection of these could be used to indicate the quality level of the sequencing.

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

To better enable the assessment of the differences in the sequences generated by the laboratories, each of the six strains in the wet-lab component was sequenced on the PacBio to obtain a closed reference genome. Initially, 10 kb template libraries were created using “10 kb DNA Template Prep Kit 1.0” from Pacific Biosciences. Subsequently, the libraries were sequenced using C2 chemistry on single-molecule real-time (SMRT) cells with a 180 min collection protocol. The data were *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. Unfortunately, the *Listeria monocytogenes* strain ST-121 (GMI16-004) was mixed up with another strain in the process of closing the genome and thus the data of this strain have been omitted the current report. The following reference genomes were generated for the PT, *C. coli* (GMI16-001): CFSAN054106, *C. jejuni* (GMI16-002): CFSAN054107, *L. monocytogenes* (GMI16-003): CFSAN054108, *L. monocytogenes* (GMI16-004): CFSAN054109, *K. pneumoniae* (GMI16-005): CFSAN054110, and *K. pneumoniae* (GMI16-006): CFSAN054111.

3.3 Distribution

On 24 October 2016, bacterial strains in agar stab cultures together with the corresponding purified and dried DNA and a welcome letter 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.

3.4 Procedure

The protocol was made available on the website allowing the PT laboratories access to all necessary information at any time (http://www.globalmicrobialidentifier.org/Workgroups/About-the-GMI-Proficiency-Test-2016). Additional relevant information was distributed by email directly to the laboratories.

The protocol presented instructions as to the handling of the received bacterial cultures and DNA.

Laboratories were requested to capture information in relation to the questions presented in the online survey of participants.

This report summarizes the results and allows for ensures full anonymity for the laboratories, as only the PT-organizers has access to the individual results.

3.5. Sequencing protocols and quality metrics

3.5.1. Online survey of the sequencing capabilities

Apart from three questions relating to the contact information of the laboratory, 40 questions were asked 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.

The laboratories submitted raw sequence files in fastq format. As part of the analysis, the reads were *de novo* assembled using the SPAdes v 3.6.1 software. Reads were aligned to reference chromosomes and plasmids using Burrows-Wheeler Aligner (BWA)-MEM algorithm with default settings. Samtools was used to filter the reads that did not map. MLST genes and sequence type were predicted using MLST tool provided by Center for Genomic Epidemiology (CGE) (https://cge.cbs.dtu.dk/services/MLST/) (1). Antimicrobial resistance genes were predicted using ResFinder tool https://cge.cbs.dtu.dk//services/ResFinder/database.php (4).

For the raw reads, the following QC metrics were calculated:

- Numbers of reads (for paired-end reads, the total numbers of reads is calculated as the sum of reads in the two files)
- Numbers of unmapped reads
- Number of reads that map to the total reference DNA (chromosome + any plasmids) using BWA
- 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
- Coverage of the reference chromosome (fraction of chromosome positions that were covered by at least one read pair).
- Coverage of the reference plasmid #1 - #3 (fraction of plasmid positions that were covered by at least one read pair).
- Depth of coverage of total DNA
- Depth of coverage of the reference chromosome
- Depth of coverage of the reference plasmid #1 - #3

For the assemblies, the following QC parameters were calculated:

- Total size of assembly (bp) (all contigs)
- Proportion (%) of size of assembly that map to the total size of DNA
- 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)

A number of samples were excluded from the analysis due to the detection of an incorrect MLST and antimicrobial resistance profile when compared to the reference strain. This affected the following samples GMI16-001-BACT/DNA and GMI16-002-BACT/DNA for user #114; and samples GMI16-005-BACT and GMI16-006-BACT for user #96.

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

Phylogenetic SNP trees were created using the pipeline; CSI phylogeny v.1.4 available from CGE. The paired-end reads were mapped to the reference genomes; using 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 10 bp between each SNP, 2) the mapping quality was more above 25, 3) the SNP quality was more than 30 and 4) 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 maximum likelihood tree using FastTree (3).

4. Results

4.1 Participation

A total of 47 laboratories responded to the pre-notification and were enrolled in the 3rd GMI PT wet-lab component. When the deadline for submitting results was reached, 46 laboratories (two laboratories split the participation in between two departments) in 22 countries had uploaded data. The following countries representing four continents provided data for at least one of the PT components (Figure 1): Australia (3 laboratories), Austria, Belgium (2 laboratories), Canada (2 laboratories), Denmark (3 laboratories), Finland, France, Germany (3 laboratories), Hong Kong, Italy (7 laboratories), Latvia, Luxembourg, Mexico, the Netherlands (3 laboratories), Poland, Portugal, Singapore (2 laboratories), Sweden (2 laboratories), Switzerland, Taiwan, the United Kingdom (2 laboratories), and the United States (6 laboratories).

4.2 Method description of the wet lab component

The time related to the handling of the PT-material, strains varied significantly between the laboratories from a few days to several weeks. The storage condition similarly ranged from -20˚C to room temperature. Variation was also observed in relation to the incubation time and temperature of the bacterial strains ranging from 4h to several days and from 35˚C to 42˚C. The laboratories also reported the DNA extraction procedures which indicated a high degree of variation among the kits being used. Among the 48 laboratories, 11 laboratories used an automatic extraction based on the following instuments, QIAsymphony SP (n = 3), Maxwell 16 (n =2), QIAcube HT (n = 2), MagNa Pure Compact, MagNa Pure LC, Chemagic Prepito-D, and BioRobot EZ1. In addition, the laboratories also reported the DNA concentration (ng/µl) and DNA quality (e.g. RIN, 260/280 ratio and/or 260/230 ratio) prior to library preparation. The DNA concentration (ng/µl) prior to library preparation was measured on Qubit (n = 36), Nanodrop (n = 6), Quant-IT on Varioskan plate reader, PicoGreen DNA Assay, Biospectomete, GloMax® 96 Microplate Luminometer, an automated plate reader using fluorescence, and others (n = 2). The DNA quality (e.g. RIN, 260/280 ratio and/or 260/230 ratio) prior to library preparation was measured on Bioanalyser (n = 3), Nanodrop (n = 18), Agilent Tape Station: Genomic DNA Screen Tape (n = 3), Qubit (n = 2), BioDrop instrument, Biospectometer, BioPhotometer plus, quantifluor kit read on POLARstar Omega plate reader, automated plate reader using absorbance, and on a gel (n = 1) in contrast to 16 laboratories which did not measure the quality. Almost all laboratories used commercial kits for library preparation and all related to the used sequencing platform, Nextera XT DNA Library Preparation Kits (n = 38), TruSeq Nano DNA HT Sample Preparation Kit (n = 2), NEBNext® Ultra™ II DNA Library Prep Kit (n = 3),

Ion Express Plus Fragment Library Preparation kit (n = 2), and ION Shear Plus Reagents kit. The genomic DNA was prepared for pair-end sequencing by 45 (93.8%) laboratories whereas two laboratories, #105 and #110 (4.2%) prepared for single-end (Ion Torrent users). One laboratory, #115 did not reply to the question. The libraries were sequenced by 35 (72.9%), seven (14.6%), and three (6.3%) laboratories using an Illumina platforms, MiSeq, NextSeq500 and Hiseq 2000/2500. Three laboratories used a Life Technology platform, two laboratories, #105 and #113 (4.2%) laboratories used the Ion Torrent PGM and one laboratory, #110 (2.1%) laboratory used the Ion

Torrent S5XL. The read length of the sequences ranged between 76 and up to 650 bp with a median at 250 bp. The reads were trimmed before upload by 10 (20.8%) laboratories using different automatic or in house tools such as FASTQ Toolkit (https://basespace.illumina.com/app), Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic), A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data (https://sourceforge.net/projects/ngopt/), trim_galore with the following settings: trim_galore --max_n 1 -o trimmed -q 20 --length 150 --paired, Torrent Server default workflow, and Btrim. Twenty-three laboratories indicated that if assembled by themselves, they would have used SPAdes. In contrast, five and four laboratories would have used Velvet and CLCbio, respectively. The remaining five employed either the A5-miseq, CGE Assembler, GS Assembler, FullSpades, patricbrc.org, or MIRA 4.0.2 assemblers.

4.3 Sequencing, wet lab – Quality markers

Out of the 46 laboratories participating in the PT trial, the following laboratories, #86, #98, #107, #112, #113, and #116 did not submit sequencing data for the *Campylobacter* GMI16-001 and GMI16-002 related to the quality metrics and parameters from both the bacterial culture and corresponding DNA. For *Listeria* GMI16-003, seven laboratories did not participate, i.e. #93, #95, #101, #103, #111, #114, and #116 in the testing of either the bacterial culture or corresponding DNA. In testing the *Klebsiella* strains, GMI16-005 and GMI16-006, the following laboratories did not participate: #71, #76, #91, #95, #101, #103, #111, #113, #115, #117, and #118 in either testing of bacterial cultures or the corresponding DNA.

The sequencing quality of all submitted data was evaluated for potential contamination or a low performance by accessing the above quality parameters.

For the *Campylobacter* genome, GMI16-001-BACT with a genome size of 1,925,274 bp, laboratory #71, #79 and #83, were considered outliers with assembled genome sizes from 3,212,811 bp (167%) (#71) to 6,391,232 bp (332%) (#79) (Figure 2 and 3). Similarly for the GMI16-001-DNA sample, laboratories #71 and #80 were considered outliers.

For the second *Campylobacter* genome, GMI16-002-BACT with a genome size of 1,898,513 bp, laboratory #83, was again considered an outlier with the highest assembled genome size of 4,203,448 bp (221%) with also laboratory #71 and #91 considered outliers with values close to those of laboratory #83 (Figure 2 and 3). Laboratory #71 was the laboratory which provided the highest assembled genome size of the GMI16-002-DNA sample with 4,313,450 bp (227%) and was considered an outlier. Laboratory #80 and #104 also reported large assembled genome size for this sample and were similarly considered outliers. The same three laboratories, #80, #83 and #93 were all considered outliers measuring the number of reads produced for both *Campylobacter* genomes, GMI16-001-BACT and GMI16-002-BACT, with laboratory #93 reaching 26,520,020 bp and 27,296,518 bp, respectively (Figure 4). For the DNA sample of the *Campylobacter* genomes GMI16-001 and GMI16-002, laboratory #93 was also considered an outlier followed by laboratory #80 for genome GMI16-002 (Figure 4). The laboratory which reported the highest amount of unmapped reads for the *Campylobacter* genomes, GMI16-001-BACT and GMI16-002-BACT were laboratory #83 with 7,588,368 bp and 8,157,147 bp, respectively and considered an outlier (Figure 5). In addition, also laboratory #115 was considered an outlier for sample GMI16-001-BACT. All of the laboratories reported data almost no unmapped reads for the DNA samples of the *Campylobacter* genomes. Some consistency in the laboratories underperforming was also observed when reporting the number of reads mapping to the reference chromosome, plasmid 1 and proportion mapping to the reference DNA sequence (Figure 6-9). Laboratory #93 reported the highest number of reads mapping to the reference chromosome for both samples types and for both *Campylobacter* genomes e.g. 24,652,082 bp for GMI16-001-BACT thus being considered as an outlier (Figure 6). Laboratory #80 (all sample types and *Campylobacter* genomes except for GMI16-001-DNA) and partly laboratory #101 (only GMI16-001-BACT) were outliers with values just above the standard deviations. All laboratories except for #83 (12%/GMI16-001-BACT, 4%/GMI16-002-BACT) and #115 (56%/GMI16-001-BACT) obtained an almost 100% in the proportion of reads mapping to the reference DNA sequence (Figure 7). Both laboratories were considered outliers. The coverage to the reference chromosome of GMI16-002-BACT was similarly low for laboratory #83 with 95% and this laboratory was considered an outlier (Figure 10). In addition, the laboratories, #69 and #97 both were considered outliers for both sample types of GMI16-001 in the coverage to the reference chromosome. The *Campylobacter* genomes only contained one plasmid (reference size 1,894,971 bp) for which laboratory #93 obtained 1,825,643 bp mapping reads (GMI16-001-BACT) and considered an outlier (Figure 9). The laboratory, #93 was similarly an outlier for both sample types and both genomes with the highest mapping reads among all laboratories. In addition to laboratory #93, also laboratories, #71, #80, #101 and #111 were considered outliers for either one or more sample types and genomes. All laboratories obtained 100% coverage for the plasmid #1 of GMI16-001 whereas the laboratory, #83 obtained a lower coverage of 97% (Figure 11). The laboratories, #80, #93, and #101 that were considered outliers compared to the other laboratories for mapping reads to the reference chromosome also sequenced the genomes deep with a sequencing depth of e.g. 1375X for GMI16-001-BACT (Figure 6, 12, and 13). This included also the depth for the chromosome and plasmid (Figure 12 and Figure 14). The total number of contigs and those above 200 bp were estimated and a number of laboratories were considered outliers including laboratory #71 for the *Campylobacter* genomes of both sample types (Figure 15 and 16). Laboratory #71, obtained 2,205; 4,138; 3,593 and 3,750 contigs for both *Campylobacter* genomes, GMI16-001 and GMI16-002 and both samples types, respectively. In addition, laboratory #79 obtained a staggering high number of contigs estimated to 6,363 for GMI16-001-BACT and was considered an outlier along with laboratories #71, #79, #80, #83, #91, #104, and #105 (Figure 15). Laboratories, #71, #79, #83, #91, #104, and #105 all obtained contigs greater than 200 bp (Figure 16). In contrast to the contigs, also the N50 and NG50 were estimated based on the submitted sequence data. The laboratories considered outliers related to the number of contigs were similarly outliers related to the N50 e.g. 1,442 bp (laboratory #79) for GMI16-001-BACT (Figure 17). This did not fully comply with NG50 where especially laboratory #105 seemed to be the most profound outlier for the *Campylobacter* genomes of both sample types (Figure 18).

For the *Listeria* genome, GMI16-003-BACT with a genome size of 2,994,410 bp, laboratory #71, #80, #86, #91, and #105, were considered outliers with assembled genome sizes from 3,219,512 bp (108%) (#105) to 4,421,076 bp (148%) (#71) (Figures 2 and 3). Similarly for the GMI16-003-DNA sample, the same laboratories as for the culture sample were considered outliers except for #86 and #91.

Laboratory #80 and #107 were both considered outliers measuring the number of reads produced for both samples types of the *Listeria* genome, GMI16-003 BACT and GMI16-003-DNA with laboratory #107 reaching 49,702,094 bp and 38,775,578 bp, respectively (Figure 4). The same laboratories, #80 and #107 were considered outliers for sequencing depth of e.g. 1554X for GMI16-003-BACT (#107) and considered outliers compared to the other laboratories (Figure 12 and 13). The laboratories which reported the highest amount of unmapped reads for the *Listeria* genomes, GMI16-003-BACT and GMI16-003-DNA were laboratory #83 and #107 with 2,640,551 bp (#107) and 527,632 bp (#83), respectively and considered outliers (Figure 5). All of the remaining laboratories reported data with almost no unmapped reads for both sample types of the *Listeria* genome. Some consistency in the laboratories underperforming was also observed when reporting the number of reads mapping to the reference chromosome and proportion mapping to the reference

DNA sequence (Figure 6 and 7). Laboratory #107 reported the highest number of reads mapping to the reference chromosome for both samples types of the *Listeria* genome e.g. 47,061,543 bp for GMI16-003-BACT, thus, was considered as an outlier (Figure 6). Laboratory #80 (both sample types) was also an outlier with values above the standard deviations. In general, the proportion of reads mapping to the reference DNA sequence were lower than 100% for most of the laboratories with the following considered as outliers, #71, #83, #91 (only GMI16-003-BACT), #105, #107 (only GMI16-003-BACT), and #113 (only GMI16-003-DNA) (Figure 7). The coverage to the reference chromosome of GMI16-003 and both sample types were quite high with almost all laboratories having coverage of 100% (Figure 10). The total number of contigs and those above 200 bp were estimated and a number of laboratories were considered outliers including laboratory #71, #80, #86, #91, and #105 for the *Listeria* genome GMI16-003-BACT (Figure 15 and 16). The same laboratories were considered outliers for the *Listeria* genome GMI16-003-DNA with the exception of the laboratories, #86 and #91. Laboratory #71, obtained 2,587 and 1,217 contigs above 200 bp for the *Listeria* genome, GMI16-003 and both samples types, respectively. In contrast to the contigs, also the N50 and NG50 were estimated based on the submitted sequence data. The laboratories considered outliers related to the N50 were #69 (DNA only), #71 (BACT only), #79 (DNA only), #80, #90 (DNA only), #91 (BACT only), #97, #105, #107 (BACT only), #110 (BACT only), #113 (DNA only), and #115 (BACT only) for GMI16-003 with e.g. 3,088 bp for GMI16-003-BACT (#105) (Figure 17). This fully complied with NG50 (Figure 18). Especially laboratory #105 seemed to be the most profound outlier for the *Listeria* genomes of both sample types.

For the *Klebsiella* genome, GMI16-005-BACT with an genome size of 5,709,497 bp, laboratory #80 and #104 were the only laboratories considered an outlier e.g. a slightly larger assembled genome size of 6,375,621 bp (112%) for laboratory #104 (Figures 2 and 3). Similarly, for the GMI16-005-DNA sample, only laboratory #80 and #105 were considered outliers but with laboratory #105 having a considerable larger assembled genome size than expected (29,882,603 bp, 523%).

For the second *Klebsiella* genome, GMI16-006-BACT with a genome size of 5,535,332 bp, laboratory #80, was considered an outlier with the highest assembled genome size of 6,894,168 bp (125%) (Figure 2 and 3). Laboratory #80 was the laboratory which again provided the highest assembled genome size of the GMI16-006-DNA sample with 5,845,637 bp (106%) and was considered an outlier. In addition, also laboratory #75, #79, #105, and #120 reported large assembled genome size for this sample and were similarly considered outliers. Laboratory #80, #86 #93, and #107 were all considered outliers measuring the number of reads produced for both samples types and both *Klebsiella* genomes with the exception of laboratory #86 (Figure 4). Laboratory #86 was only considered an outlier for GMI16-005 BACT. Among the laboratories considered outliers, Laboratory #107 produced the highest amount of reads reaching 14,912,020 bp (GMI16-005-DNA) and 32,168,468 bp (GMI16-006), respectively (Figure 4). The same laboratories, #80 and #107 were considered outliers for sequencing depth of the chromosome, total DNA, and plasmids e.g. 257X of total DNA for GMI16-005-DNA and 581X of the chromosome for GMI16-006-DNA (#107) compared to the other laboratories (Figure 12-13 and 19-20). The laboratory which reported the highest amount of unmapped reads for the *Klebsiella* genomes was laboratory #93 with 3,986,353 bp (GMI15-006-BACT) and was considered an outlier (Figure 5). In addition, also laboratory #83, #86, #93, and #107 were considered outliers for sample GMI16-006 of both sample types. In general, almost all of the remaining laboratories reported data without unmapped reads (Figure 5). Some consistency in the laboratories underperforming was also observed when reporting the number of reads mapping to the reference chromosome, reads mapping to plasmid 1-3, and proportion mapping to the reference DNA sequence and plasmids (Figure 6-9, Figure 19 and 20). Laboratory #107 reported the highest amount of reads mapping to the reference chromosome for both samples types and for both *Klebsiella* genomes e.g. 30,898,875 bp for GMI16-006-DNA thus, being considered as an outlier (Figure 6). In addition, Laboratory #68, #80, #86, and #93 were outliers with values just above the standard deviations for at least one sample type and one of the *Klebsiella* genomes. All laboratories except for #75 (90%/GMI16-006-DNA), #83 (93%/GMI16-005-BACT, 94%/GMI16-005-DNA, 37%/GMI16-006-BACT, and 90%/GMI16-006-DNA), #93 (1%/GMI16-005-BACT, 64%/GMI16-006-BACT, 95%/GMI16-006-DNA), and #107 (2%/GMI16-006-BACT) obtained an almost 100% in the proportion of reads mapping to the reference DNA sequence (Figure 8). The coverage to the reference chromosome of GMI16-005-BACT was similarly low for laboratory #93 with 76% and considered an outlier (Figure 10). In addition, the laboratory, #107 was similarly considered an outlier for GMI16-006-BACT (96%) in the coverage to the reference chromosome. The *Klebsiella* genomes contained up to three plasmids (reference GMI16-005 plasmid 1 size 227,967 bp, reference GMI16-005 plasmid 2 size 43,380 bp, reference GMI16-006 plasmid 1 size 116,768 bp, reference GMI16-006 plasmid 2 size 100,222 bp, and reference GMI16-006 plasmid 3 size 38,695 bp) for which laboratory #93 and #107 were considered outliers due to a lower number of reads and coverage of the plasmids i.e. 3,964 bp/78%

GMI16-005-BACT plasmid 1 (#93), 840 bp/84% GMI16-006-BACT plasmid 1 (#107), 675 bp/83% GMI16-006-BACT plasmid 2 (#107), and 427 bp/88% GMI16-006-BACT plasmid 3 (#107) (Figure 9, 11, and 19-22) and also due to a high depth of coverage for the three plasmids (Figure 14, 23 and 24). The total number of contigs and those above 200 bp were estimated and a number of laboratories were considered outliers including laboratories #75, #80, #104 (only BACT), and #105 for both sample types of *Klebsiella* genome, GMI16-005 (Figure 15 and 16). Laboratory #105, obtained 846 and 800 as well as 172,512 and 50,844 contigs and those above 200 bp for the BACT and DNA samples, respectively of *Klebsiella* genome, GMI16-005.

Some of the same laboratories produced a high number of contigs of which only a few were above 200 bp for GMI16-006 of both sample types and were considered outliers. The outliers included the following laboratories, #75, #79, #80, #83, #93, and #105 of which #80 provided most contigs of which only a few were above 200 bp for the BACT sample (13,192/430). Only laboratory #105 was considered an outlier estimating the N50 and NG50 for the *Klebsiella* genomes (Figure 17 and 18). This only accounted for GMI16-005-DNA i.e. N50: 349/NG50: 194.

4.4 Sequencing, wet lab – Phylogeny

A few SNPs were observed between the laboratories submitted genome data and the reference genome which included the following laboratories and number of SNPs, #83 (3 SNP in GMI16-001-BACT and 1 SNP in GMI16-002-BACT), #86 (1 SNP in GMI16-003-BACT and GMI16-003-DNA), respectively, and #90 (1 SNP in GMI16-005-BACT). In contrast, a total of 13,486 SNPs was observed between GMI16-004-BACT and the reference genome by laboratory #120. A similar high number of SNPs (1,792) was observed between GMI16-005-BACT and the reference genome by laboratory #96. No SNP differences were observed for the *Klebsiella* genome, GMI16-006 between the laboratories submitted genome data and the reference genome.

4.5 Sequencing, wet lab – Quality control thresholds

One of the purposes of this PT was to propose tentative quality control thresholds for satisfactory performance conducting the whole genome sequencing. Evaluating the QC parameters, only the parameters that would be available when conduction the QC on a daily basis were evaluated. These did not include the parameters associated with the closed genomes of the test strains.

We observed that some of these QC parameters would be depended on what genus was tested why some of the threshold needs to be defined on this basis. We determined the tentative quality control thresholds on the basis of the calculated double standard deviations, either the upper or lower level. The defined outliers performing poorly, all produced data above or below the level of the calculated double standard deviations. The upper double standard deviations for the size of the assembled genomes were at 1,957,767 bp, 1,960,095 bp, 5,799,052 bp, and 3,287,156 bp for *C. coli, C. jejuni*, *L. monocytogenes*, and *K. pneumoniae*, respectively. The upper double standard deviations for the number of contigs were at 203, 240, 276, and 308 for *C. coli, C. jejuni*, *L. monocytogenes*, and *K. pneumoniae,* respectively. Similarly, lower double standard deviations for N50 were at 88,184 bp, 66,566 bp, 44,489 bp, and 10,301 bp, for *C. coli, C. jejuni*, *L. monocytogenes*, and *K. pneumoniae*, respectively.

4.6 Sequencing, wet lab – MLST, and antimicrobial resistance genes

For *Campylobacter*; GMI16-001 the expected MLST was ST7426 which were identified by almost all of the laboratories except for laboratory #75 which identified GMI16-001-DNA as ST7039 why the incorrect MLSTs. Some laboratories did not report MLST data (own tool) but these were provided by PT-organizer (CGE tool) and identified the correct MLST.

The *Campylobacter*; GMI16-001 was pan-susceptible therefore no AMR genes were expected. For four laboratories, #79, #83, #114, and #118 out of 39 laboratories, deviating results were identified when analyzing the data using the CGE reference tool. Laboratory #114 reported resistance data for the bacterial sample matching the profile of GMI16-002 probably due to mix up of the test material.

The MLST ST6238 was expected for the *Campylobacter* strain; GMI16-002. This was similarly reported correctly by almost all of the participants except for laboratory #80 and #83. The CGE reference tool could not identify the ST for GMI16-002-DNA using the data provided by laboratory #80. The laboratory managed, however, to report the correct MLST using own tools. During this study all the reads were considered in the assembly process, however, trimming the low quality ends could potentially improve the assembly and it is possible to obtain the correct MLST Sequence Type. Laboratory #83 reported the incorrect MLST, ST32 for sample, GMI16-002-BACT.

A very high degree of concordance was observed between the reported AMR genes detected by own tools and the CGE reference tool and between culture (n = 35) and DNA (N = 40) samples. Some of the AMR genes, were determined “like” which indicate that the homology to the reference genes were less than 100% which is often seen due to minute sequencing errors or the tool settings

(% identity and coverage). The gene *aph*(2”)-like was reported by a number of laboratories. In contrast, the CGE tool did not detect this specific gene. This does not mean that the gene is not present but that different methods might have been applied. Most likely, the commandline version of the CGE ResFinder tool did not pick up this gene due to a higher threshold in homology than the one used by the other participants or due to the fact that in this study reads were not trimmed before the assembly and the quality of the assembly might be lower. A few laboratories, #67, #82, #83, and #84 reported a few additional resistance genes not detected the CGE reference tool e.g. laboratory #82 reported chromosomal point mutations which are not yet included the used version of the CGE ResFinder tool why this very well could be true. Running the commandline version of the CGE ResFinder tool for the genome of *Campylobacter* strain; GMI16-002 submitted by laboratory #114 showed resistance genes that do not match any of the expected profiles of the PT strains.

Almost all laboratories reported the correct MLST, ST2 for the *Listeria* strains GMI16-003 except for laboratory #84 which for both sample types using own tools reported ST512. This was a clear misreporting of the MLST as the reported MLST corresponded to GMI16-005 but still count as a mistake.

The *Listeria* strain GMI16-003 was pan-susceptible and only a few laboratories, #70, #82, #83, and #92 either reported a few resistance genes or a few resistance genes were identified using the CGE reference tool. This included the identification of the genes: *aac*(6')-Ib and *aac*(6')-Ib-cr in the bacterial sample from laboratory #82. This unexpected finding can potentially be explained by a partial contamination with GMI16-005 harbouring the same genes.

Thirty-four laboratories tested the *Klebsiella* strains, GMI16-005. The commandline version of the CGE MLSTFinder tool was used to test the submitted genome, GMI16-005-DNA laboratory #114 which did not submit own data. In all cases, the laboratories managed to identify the correct and expected MLST ST-512 except for two cases of misreporting. Laboratory #72 reported ST15 which correspond to GMI16-006 and laboratory #84 reported ST2 linking to GMI16-003. The same laboratories were involved in testing the *Klebsiella* strains, GMI16-006. The expected MLST was ST15 which were found by all laboratories except for those laboratories misreporting (#72 and #84) as previous mentioned.

Both *Klebsiella* strains were multidrug resistant (MDR) harbouring a number of AMR genes. *Klebsiella* strains, GMI16-005 were found to contain the following genes, *aad*A2, *aac*(6')-lb), *bla*$_{\text{TEM-1A}}$, *bla*$_{\text{KPC-3}}$, *bla*$_{\text{OXA-9}}$, *bla*$_{\text{SHV-11}}$, *oqx*A, *oqx*B, *aac*(6')lb-cr, *fos*A, *mph*(A), *cat*A1, *sul*1, and *dfr*A12. Most of the genes were identified by all laboratories and by both own and CGE tools indicated by a high concordance. Several of the laboratories report the genes being with a lower homology than the reference gene indicated by being determined “like”. A very few number of laboratories reported different variants of the expected resistance gene e.g. *aad*A1 instead of the expected *aad*A2. In addition, some laboratories reported the gene but without the variant number e.g. *bla*$_{\text{SHV}}$ instead of the expected *bla*$_{\text{SHV-11}}$. These deviations were all considered minor mistakes and not related to the sequencing quality but detection of the resistance genes or by the analytic method. The mutation, aac(6')lb-cr was not identified by laboratory #82 using own tools for both the culture and DNA in contrast to the CGE tool. The laboratory, however, identified the presence of the gene, aac(6')lb as all did. The aac(6')lb-cr gene in the CGE tool has mutations compared to aac(6')lb but these mutations do not lead to the amino acid changes needed for fluoroquinolone resistance. Therefore, the aac(6')lb-cr in the CGE tool should not have “-cr” in the name. Similarly, the CGE tool was not able to detect neither the aac(6')lb nor the mutation aac(6')lb-cr in GMI16-005-DNA for laboratory #104 indicating a potentially truncated gene. Almost all of the laboratories identified the gene, fosA in a “like” version. The commandline version of the CGE ResFinder tool did not pick up this gene most likely due to a higher threshold in homology than used by the laboratories.

Two laboratories, #92 and #96 reported more genes as expected for the culture sample using its own tool compared to the reference CGE tool. This was pronounced for laboratory #96 and could suggest laboratory contaminations. For laboratory #92, the additional detected genes might be related to improper settings for the used tool, uncertain how to interpret the genotypic data, used a tool of poor performance or were similarly affected by laboratory contaminations.

The *Klebsiella* strains, GMI16-006 contained the following genes, *aad*A1, *aac*(6')-Ib, *aac*(3)-Iid, *aph*(3')-Via, *str*A, *str*B, *bla*$_{\text{NDM-1}}$, *bla*$_{\text{OXA-9}}$, *bla*$_{\text{CTX-M-15}}$, *bla*$_{\text{SHV-12}}$/*bla*$_{\text{SHV-28}}$, *bla*$_{\text{TEM-1b}}$, *qnr*S1, *oqx*B, *oqx*A, *aac*(6')Ib-cr, *sul*2, *tet*(D), *dfr*A14, and *fos*A. The concordance was very high between the laboratories testing the strain GMI16-006. Inconsistences in detection of the gene *bla*$_{\text{SHV}}$ genes were observed. Some laboratories reported *bla*$_{\text{SHV-28}}$ and others *bla*$_{\text{SHV-12}}$, *bla*$_{\text{SHV-5}}$, and *bla*$_{\text{SHV-1}}$. Consistency, however, between “own” and CGE data was seen. Almost all of the laboratories identified the gene, fosA in a “like” version. The commandline version of the CGE ResFinder tool did not pick up this gene most likely due to a higher threshold in homology than used by the laboratories. Several laboratories, #68, #83, #84, #92, #93, #96, #99, and #106 reported the detection of the *bla*$_{\text{LEN}}$ genes. In addition, laboratory #92 reported all AMR genes without variant identification.

5. Discussion

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 in reporting the data for laboratories #72 and #84 mix up of the test genomes. The MLST was correct, however, for all PT strains when re-analysed using the CGE reference method. In the initial phase of the analysis, it was clear that a few laboratories, #96 and #114 mixed up the cultures/DNA or the genomes thus submitting genomes with the wrong name. These data have been omitted from further analysis.

Most of the submitted AMR genes were in concordance with the expected results. Some deviations, however, were observed and could be explained by either the bioinformatic tools used to identify AMR genes or contaminations. Some deviations appear to be related to using in-house tools compared to the CGE reference tool data. The plausible reasons for the deviations might the use of either a higher threshold settings of "own tool" accidentally ignoring some genes, incorrect interpretation of the AMR data provided by the "own tool", an inadequate "own tool", an "own tool" not recently updated or different assembly tool in relation to mapping the AMR genes. This could explain the profound deviations by laboratories identifying the AMR genes, *aph*(2")-If and *aph*(3')-III using as well as *aac*(6')-*Ib* using "own tool" compared to the CGE reference tool data, *aph*(3')-III and *aac*(6')-*Ib* and *aac*(6')-*Ib-cr*, respectively.

A number of laboratories identified a few of AMR genes not present in the genomes based on the CGE reference tool data. This was likely due to contaminations but in a degree which did not impact the quality control assessment with none of the laboratories indicated as an outlier.

One of the objectives for the GMI PT was to assess a range of quality markers to evaluate the performance by the participating laboratories. Overall, the PT test showed that most laboratories perform satisfactory with the exception of a few laboratories which seemed to have some challenges and designated as outliers in several related to several of the quality parameters. The laboratories performing poorly included #71, #75, #79, #80, #83, #86, #91, #93, #104, #105, #107, #115, and #120. In general, the laboratories produced either too large sizes of the expected genomes when assembled by the CGE reference assembly tool, too many reads including unmapped reads, low percentage of mapping reads to the references, a low coverage, a high number of contigs, and a low N50. In addition, a few of the laboratories also deviated with SNPs, #83, #86, #90, and #120 compared to the reference genome or mismatch with the expected AMR profiles, #79, #83, and #84 which both indicated laboratory contaminations. Thus, in general the poor performance was either a result of laboratory contamination with either cultures or DNA or issues related to the actual sequencing which could be suspected as the poor performance was equally distributed between sample types, culture and DNA.

A few of the poorly performing laboratories, #71, #80, #93, and #107, provided a huge depth of coverage (X) which could explain the large abundance of reads complicating the assembly. Most of the laboratories used high-throughput platforms such as Hiseq's, #93 and #107 and NextSeq, #80 from Illumina. In addition, a huge depth of coverage has been considered "wasting money".

Evaluating the QC parameters of this PT, the performance of the whole genome sequencing could be explained by a number of the parameters. Unfortunately, it is not likely that all of the parameters would be available when conduction the QC on a daily basis due to the absence of closed genomes of the reference test strains to estimate e.g. the number reads mapping to the reference genome, proportion mapping to the reference genome etc. It appeared, however, that a few number of QC parameters could be useful for evaluating the whole genome sequencing results such as the size of the genome, number of total reads, number of contigs, and N50. Some of these QC parameters would be depended on what genus being tested why the threshold needs to be defined on this basis. We recommend the following tentative quality control thresholds for good performance for the three genera included this PT: The size of the assembled genome should not exceed 1,960,000 bp, 1,960,000 bp, 5,800,000 bp, and 3,300,000 bp for *C. coli*, *C. jejuni*, *L. monocytogenes*, and *K. pneumoniae*, respectively. The number of contigs should not exceed 200, 250, 300, and 300 for *C. coli, C. jejuni*, *L. monocytogenes*, and *K. pneumoniae*, respectively. Similarly, the N50 should not be lower than 85,000 bp, 65,000 bp, 45,000 bp, and 10,000 bp, for *C. coli, C. jejuni*, *L. monocytogenes*, and *K. pneumoniae*, respectively.

6. Conclusions

The GMI PT 2016 provided useful data on concordance of typical analyses of bacterial genomes for the identification of epidemiological markers, MLST and AMR genes. It has also revealed critical importance of the ongoing quality assessment on the performance of dry lab and wet lab processes in whole genome sequencing. Most of the participants were able to identify the correct MLST with minor problems related to submission of the data mixing up the result of one genome with other genomes. The identification of the AMR genes provided a bit more information related to quality. Similarly, large discrepancies observed was a result of the same issue mixing up the data whereas minor changes related to the expected AMR profile were more related to sequence quality. The sequencing QC parameters were clearly better in assessing the quality of the whole genome sequencing. We identified 13 (28%) laboratories including #71, #75, #79, #80, #83, #86, #91, #93, #104, #105, #107, #115, and #120 out of 46 which performed unsatisfactory and determined outliers for one or more of the QC parameters. The poor performance was either a results of laboratory contamination of the culture and/or DNA or the sequencing itself which also appears to be somehow related to specific platforms. Tentative QC thresholds were determined for a number of parameters not depended on the availability of closed genomes of the test strains. The data suggested that these parameters being based on genus levels due to the variation across the genus.

7. Acknowledgement

The GMI PT 2016 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 PT 2016 was supported by the World Health Organization, FDA GenomeTrakr and Microbiologics®.

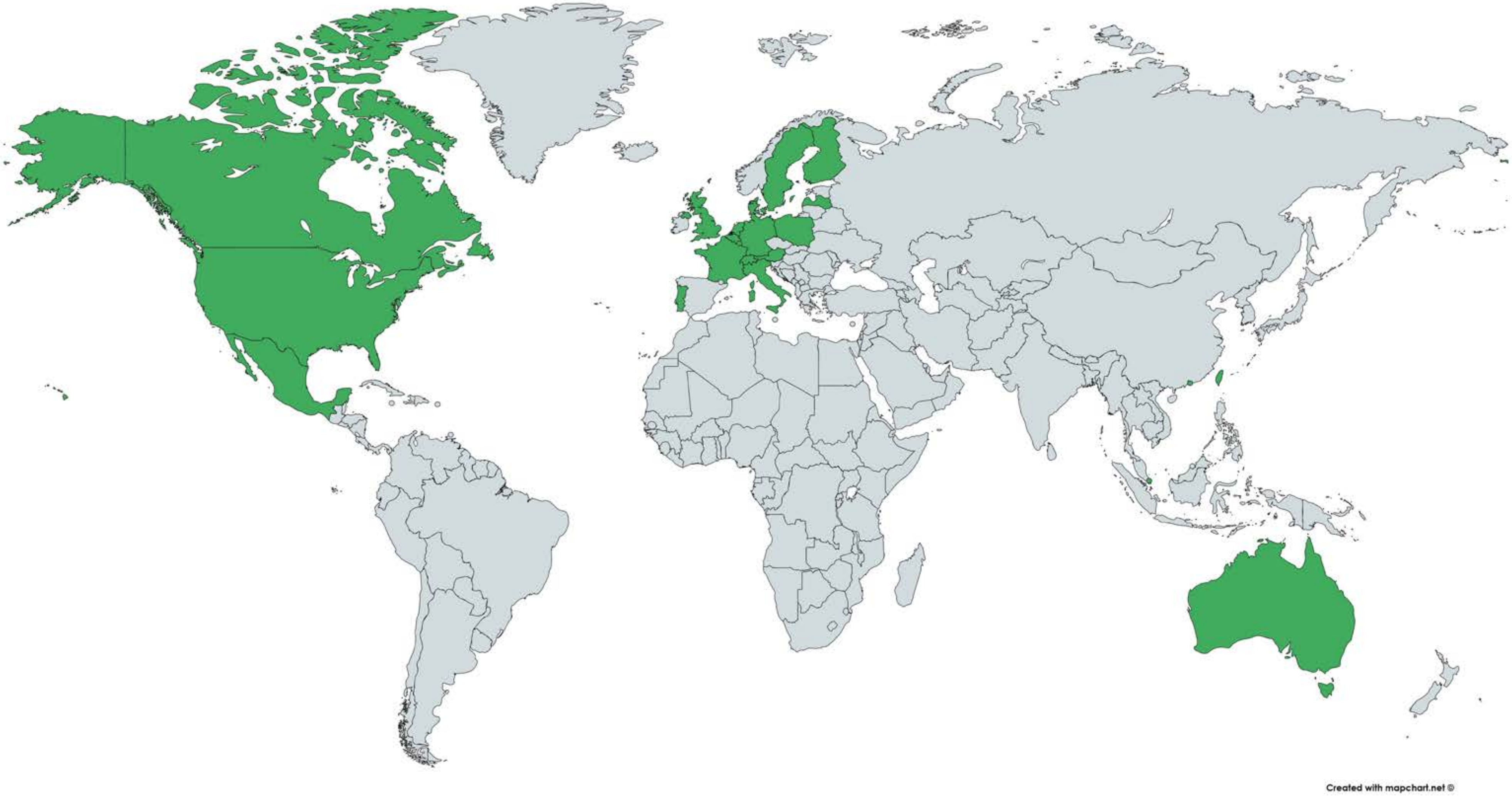
Reference List

1. **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.

2. **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.

3. **Price, M. N., P. S. Dehal, and A. P. Arkin**. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS.One. **5**:e9490.

4. **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.



Legend:
Countries marked in green participated in the wet-lab component of the GMI PT 2017

Legend to the following box plot figures (Figure 2 to 24):

Colors of boxplot points (dots) show the similarity to AMR profile, compared to the profiles identified in the reference genome.
Color scheme is as follows:

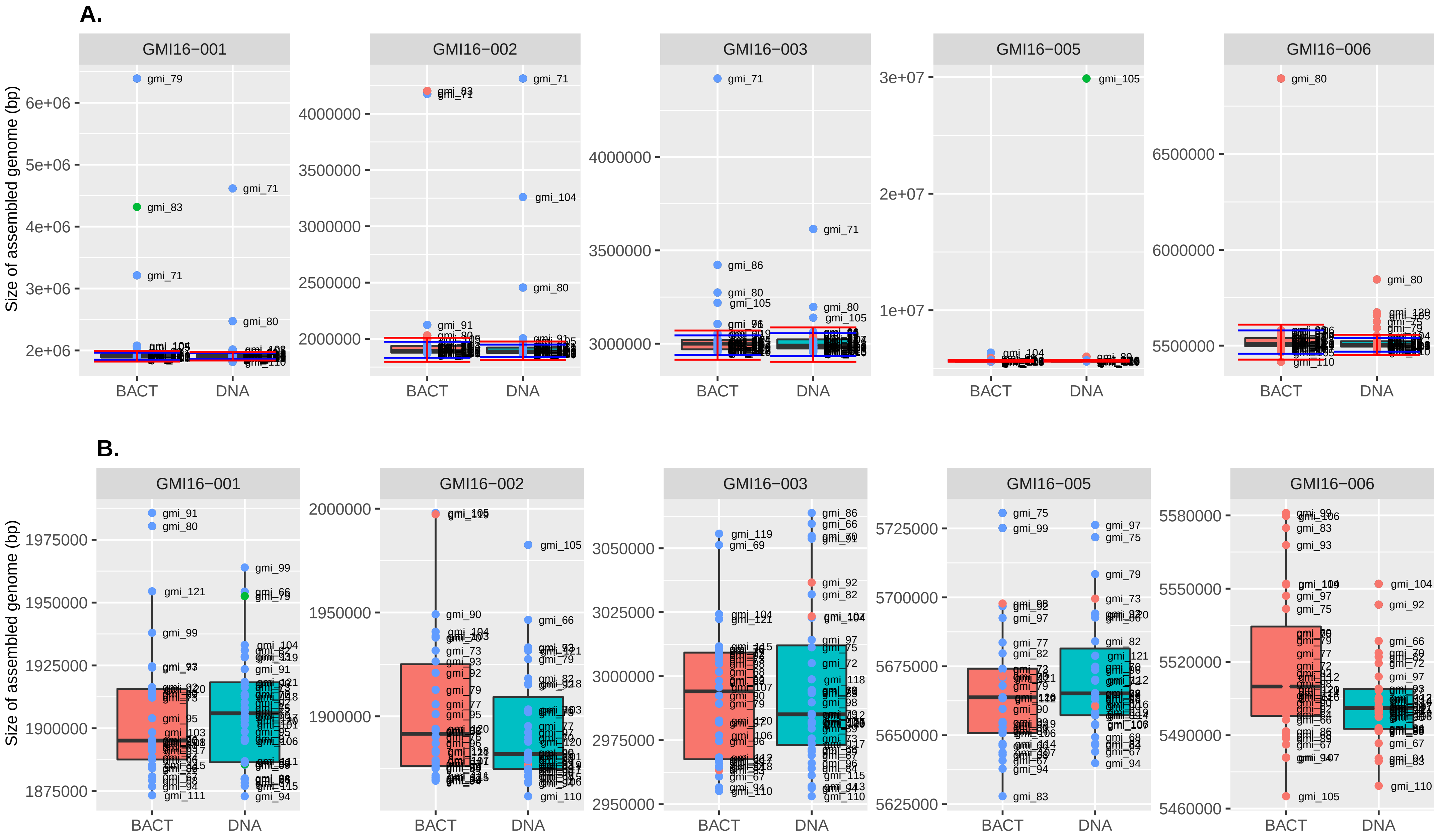
- BLUE: AMR profiles fully matches reference profile
- GREEN: AMR scheme in the participant sample is differs completely from the reference genomes
- RED: most the genes are identified correctly, but participant strain might miss one or two genes or might have identified additional genes.

Plot A shows results for all the available samples (excluding the ones with wrong MLST sequence type).
Red and blue lines indicate ± 2 and ± 3 standard deviations, respectively.

Plot B is an extract from plot A and shows the interquartile ranges of the distribution from plot A.

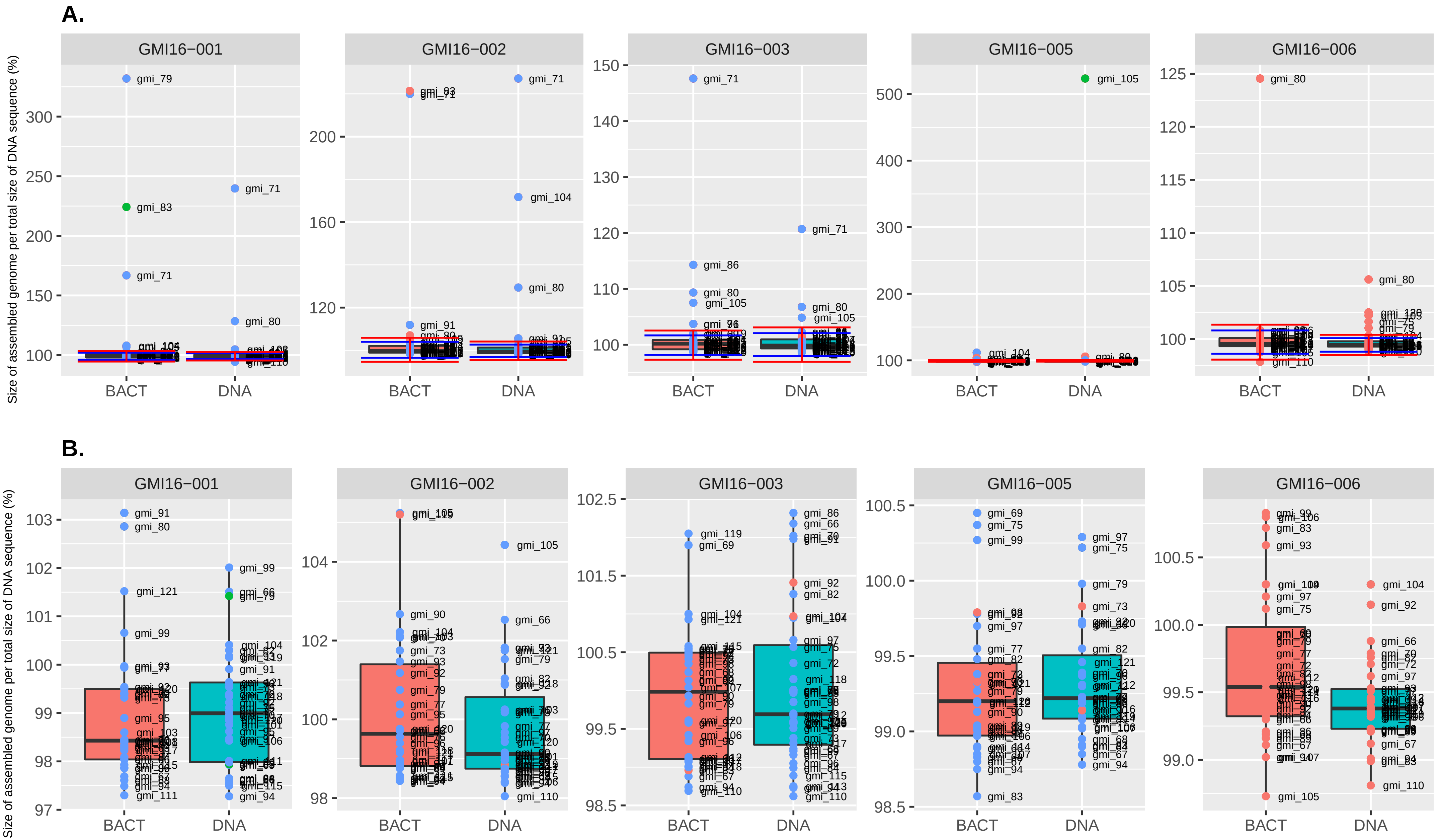
Size of assembled genome

GMI Proficiency Test 2016 Report - Figure 2



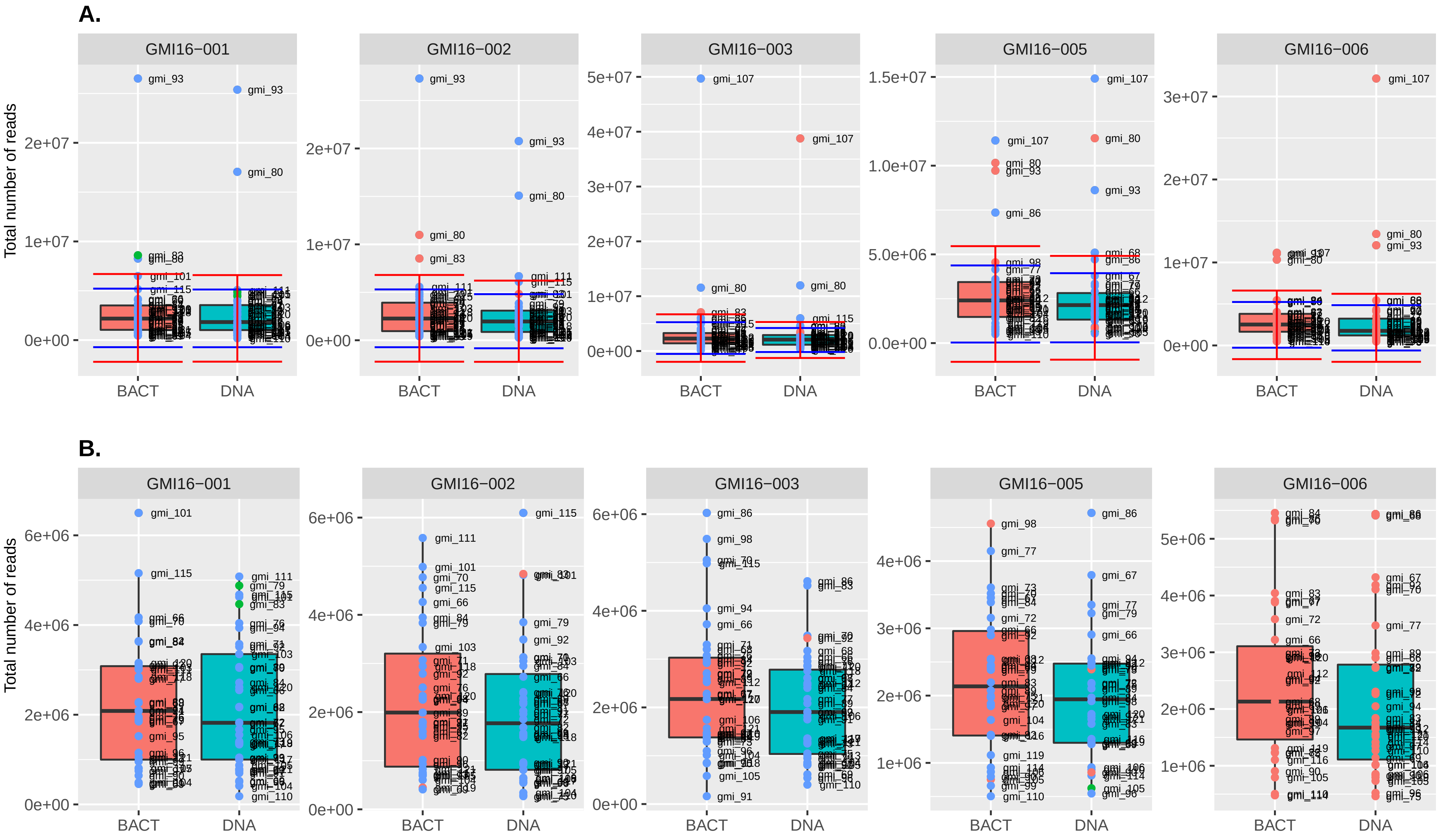
Size of assembled genome per total size of DNA sequence

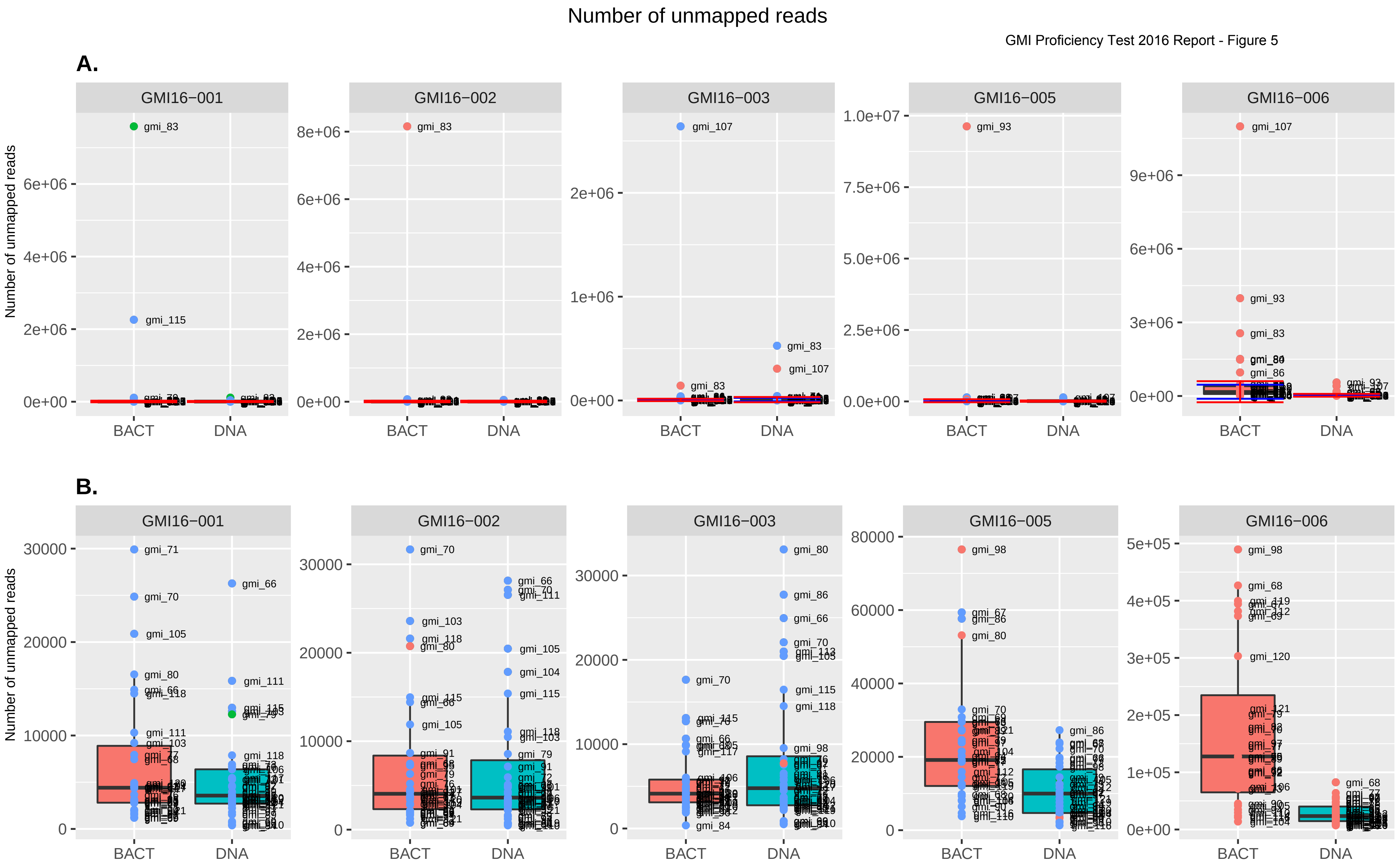
GMI Proficiency Test 2016 Report - Figure 3



Total number of reads

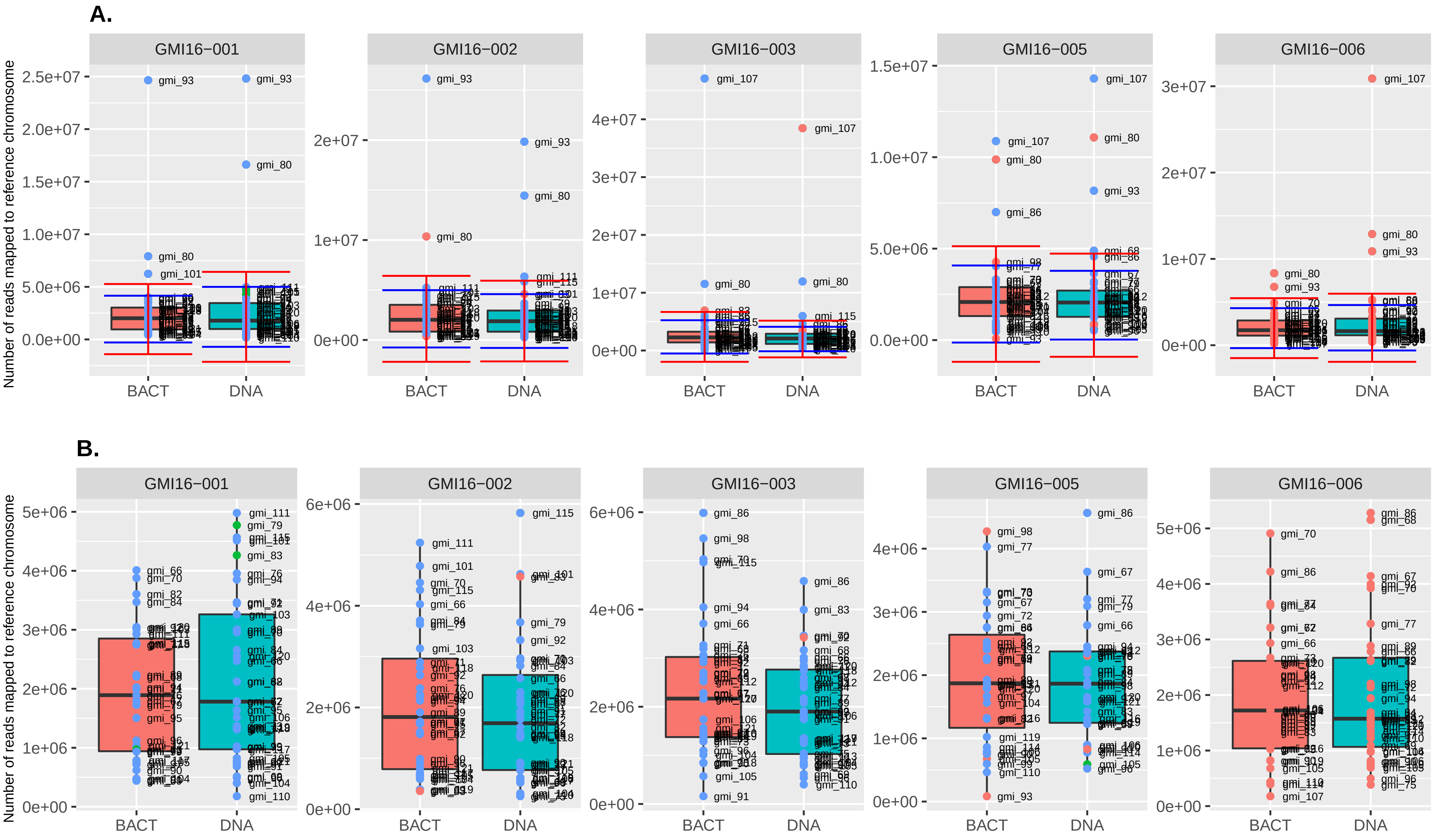
GMI Proficiency Test 2016 Report - Figure 4




Number of unmapped reads
GMI Proficiency Test 2016 Report - Figure 5
A.
GMI16−001
GMI16−002
GMI16−003
GMI16−005
GMI16−006
Number of unmapped reads
0e+00
2e+06
4e+06
6e+06
8e+06
1e+06
3e+06
9e+06
0.0e+00
2.5e+06
5.0e+06
7.5e+06
1.0e+07
BACT
DNA
gmi_83
gmi_115
gmi_107
gmi_93
gmi_86
B.
0
10000
20000
30000
40000
60000
80000
1e+05
2e+05
3e+05
4e+05
5e+05
gmi_71
gmi_70
gmi_105
gmi_80
gmi_66
gmi_118
gmi_111
gmi_103
gmi_115
gmi_104
gmi_98
gmi_86
gmi_67
gmi_120
gmi_68
gmi_84
gmi_79
gmi_91

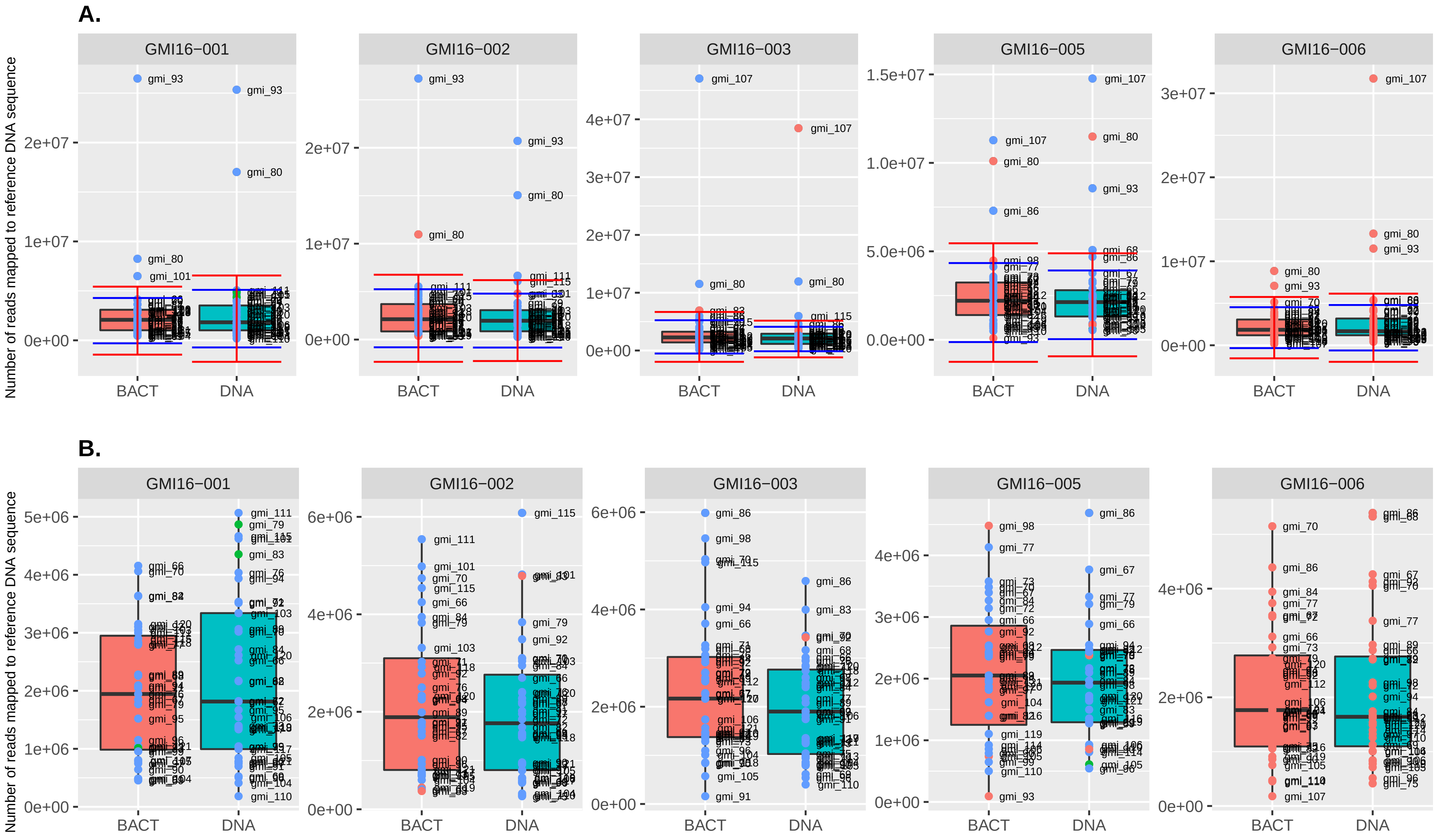
Number of reads mapped to reference chromosome

GMI Proficiency Test 2016 Report - Figure 6



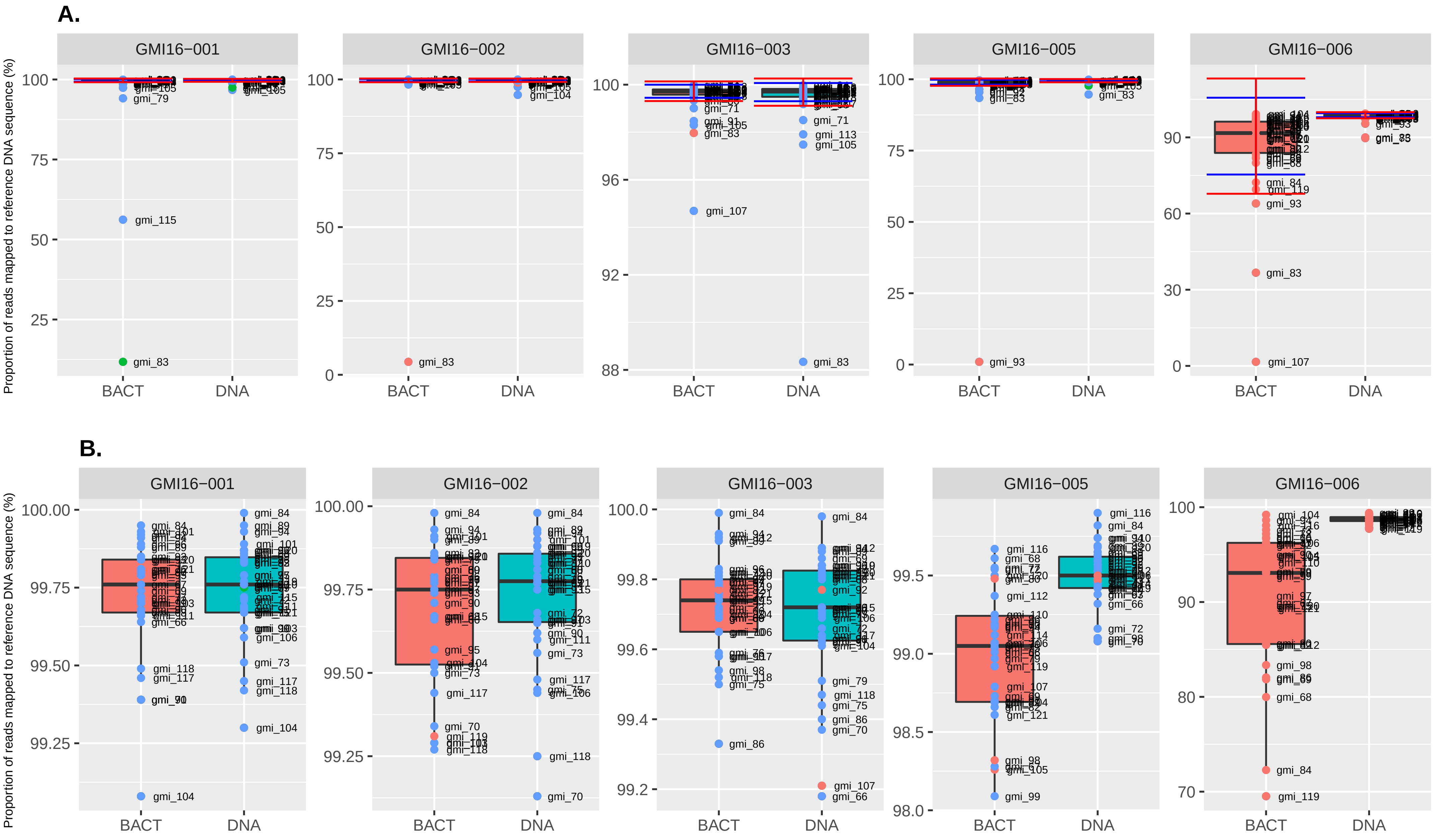
Number of reads mapped to reference DNA sequence

GMI Proficiency Test 2016 Report - Figure 7



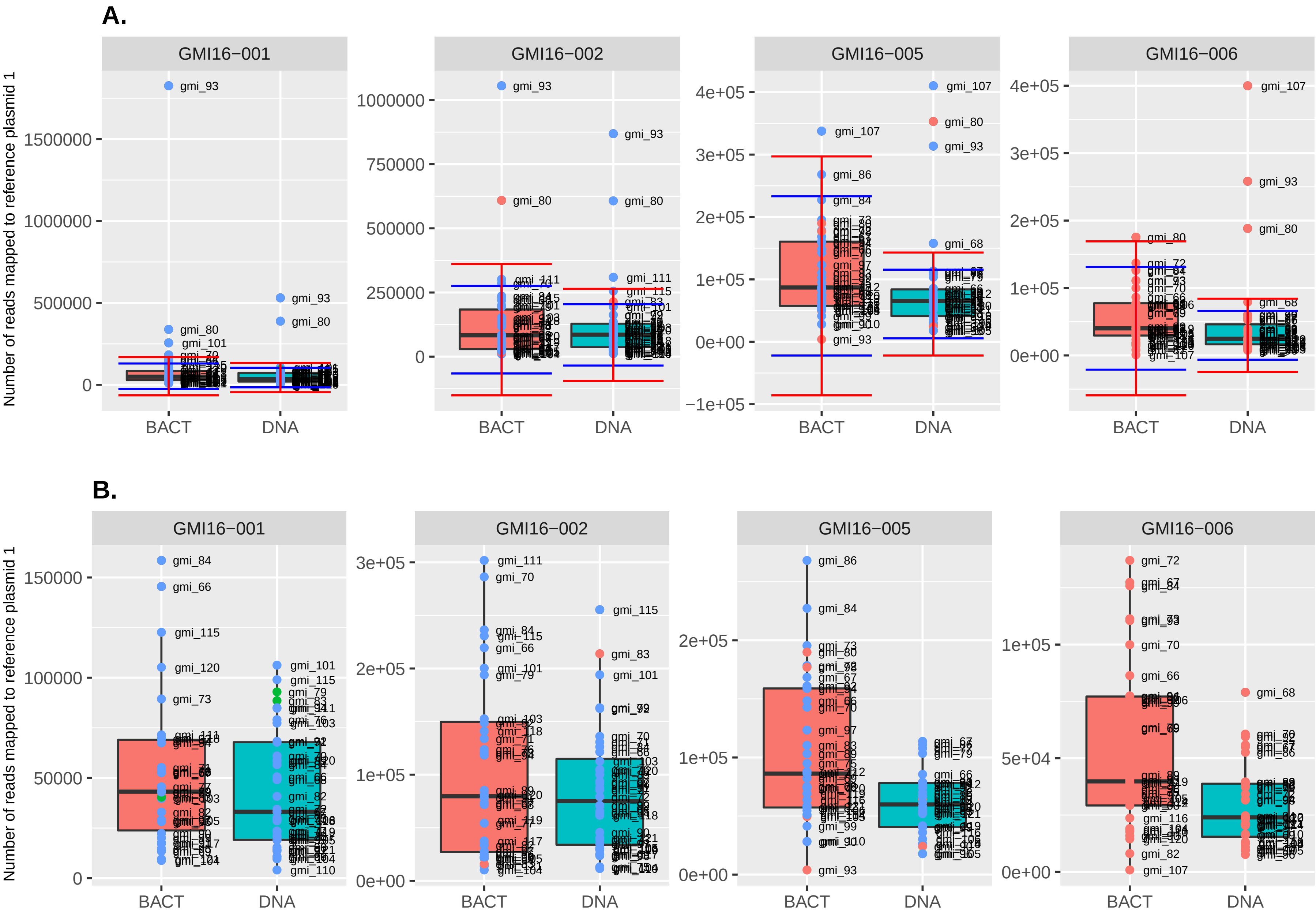
Proportion of reads mapped to reference DNA sequence

GMI Proficiency Test 2016 Report - Figure 8



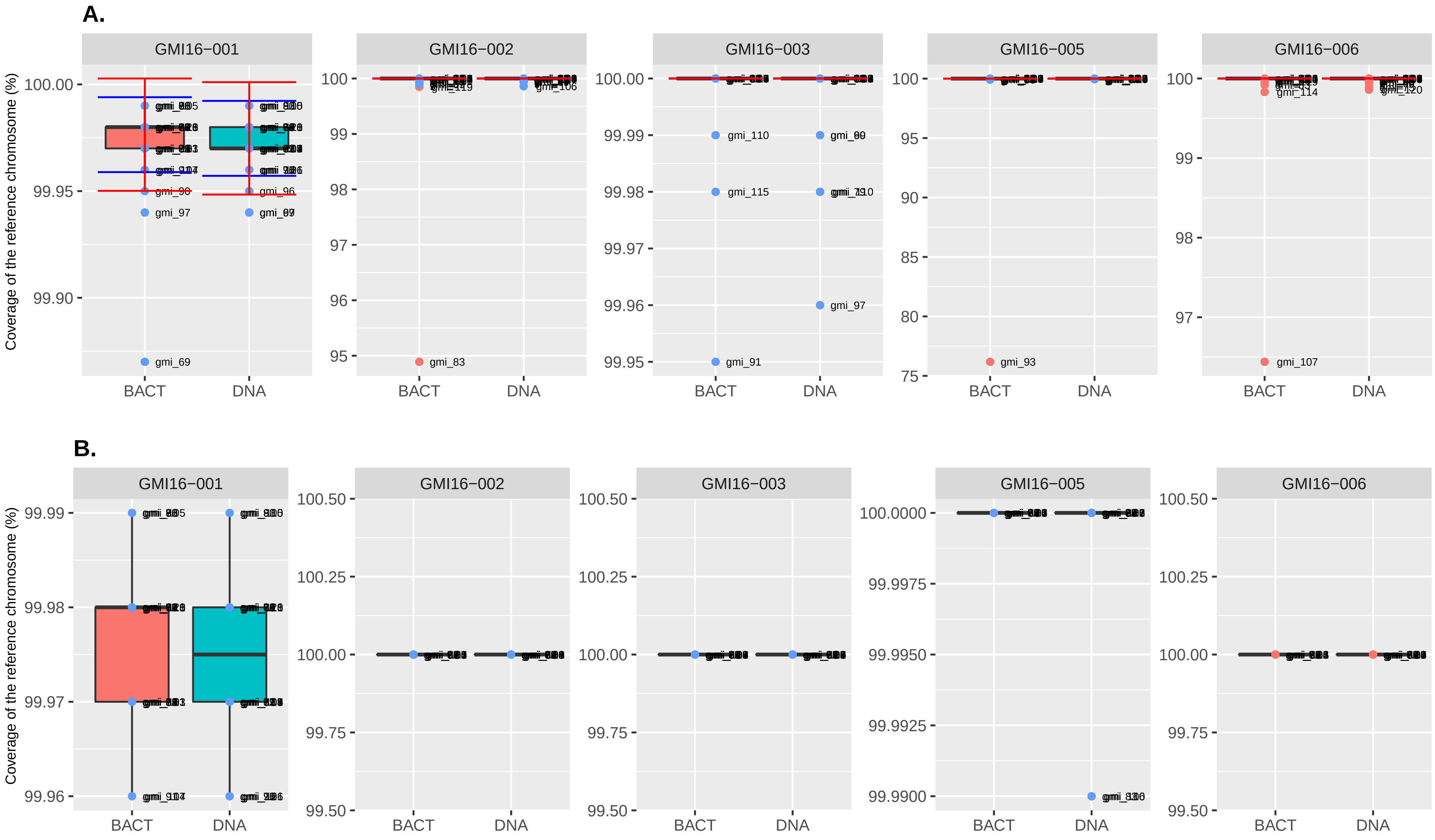
Number of reads mapped to reference plasmid 1

GMI Proficiency Test 2016 Report - Figure 9



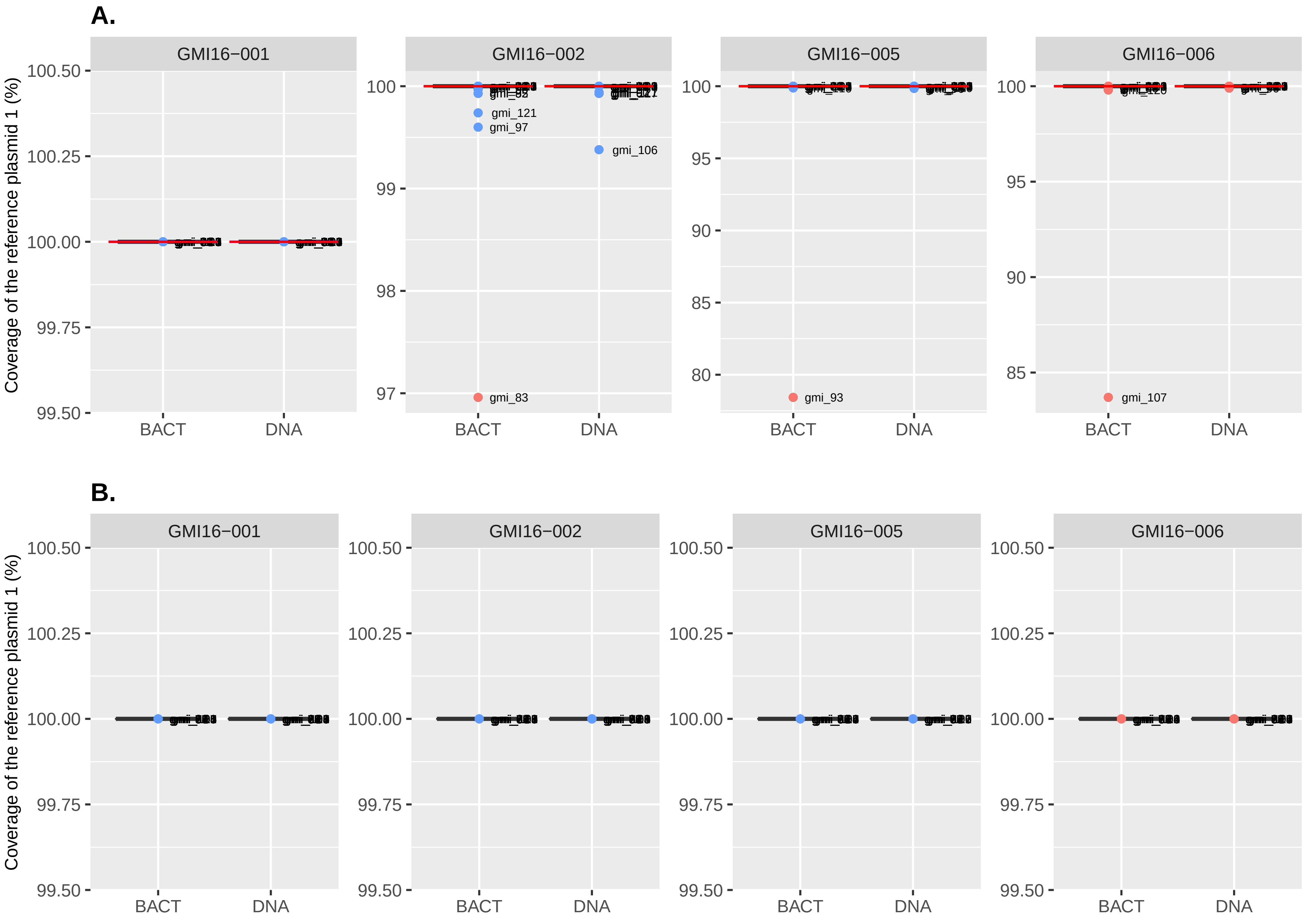
Coverage of the reference chromosome

GMI Proficiency Test 2016 Report - Figure 10



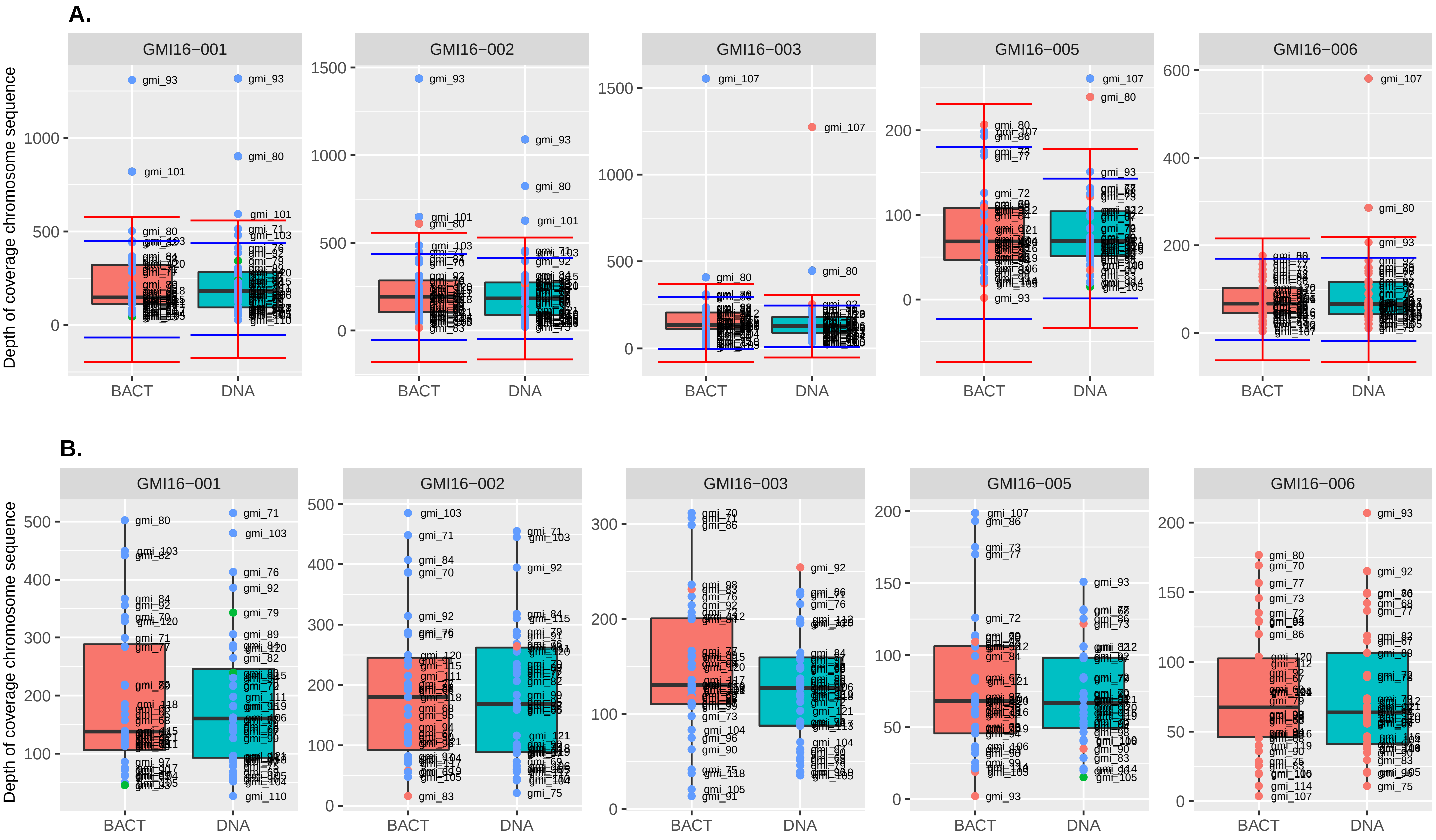
Coverage of the reference plasmid 1

GMI Proficiency Test 2016 Report - Figure 11



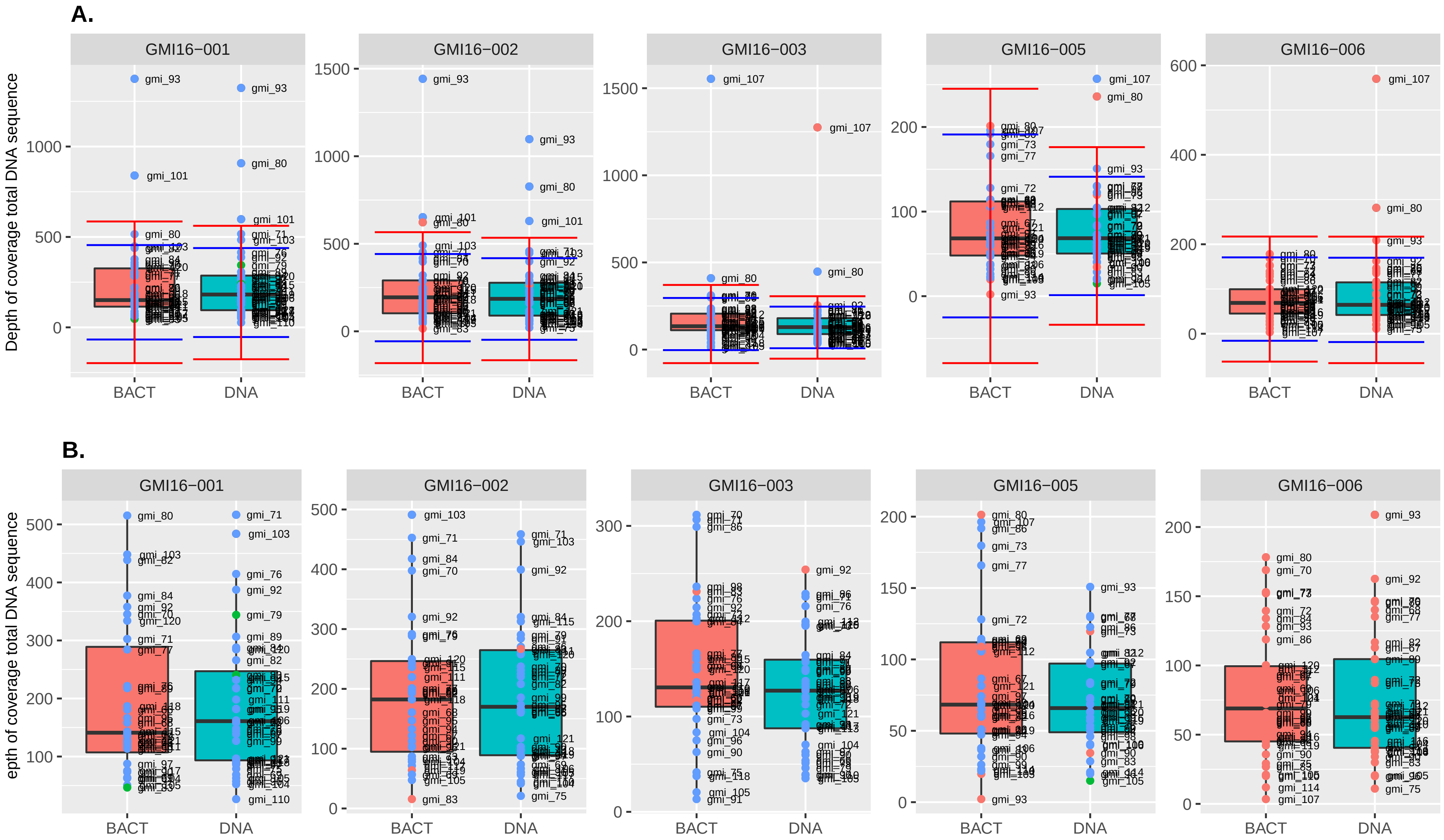
Depth of coverage chromosome sequence

GMI Proficiency Test 2016 Report - Figure 12



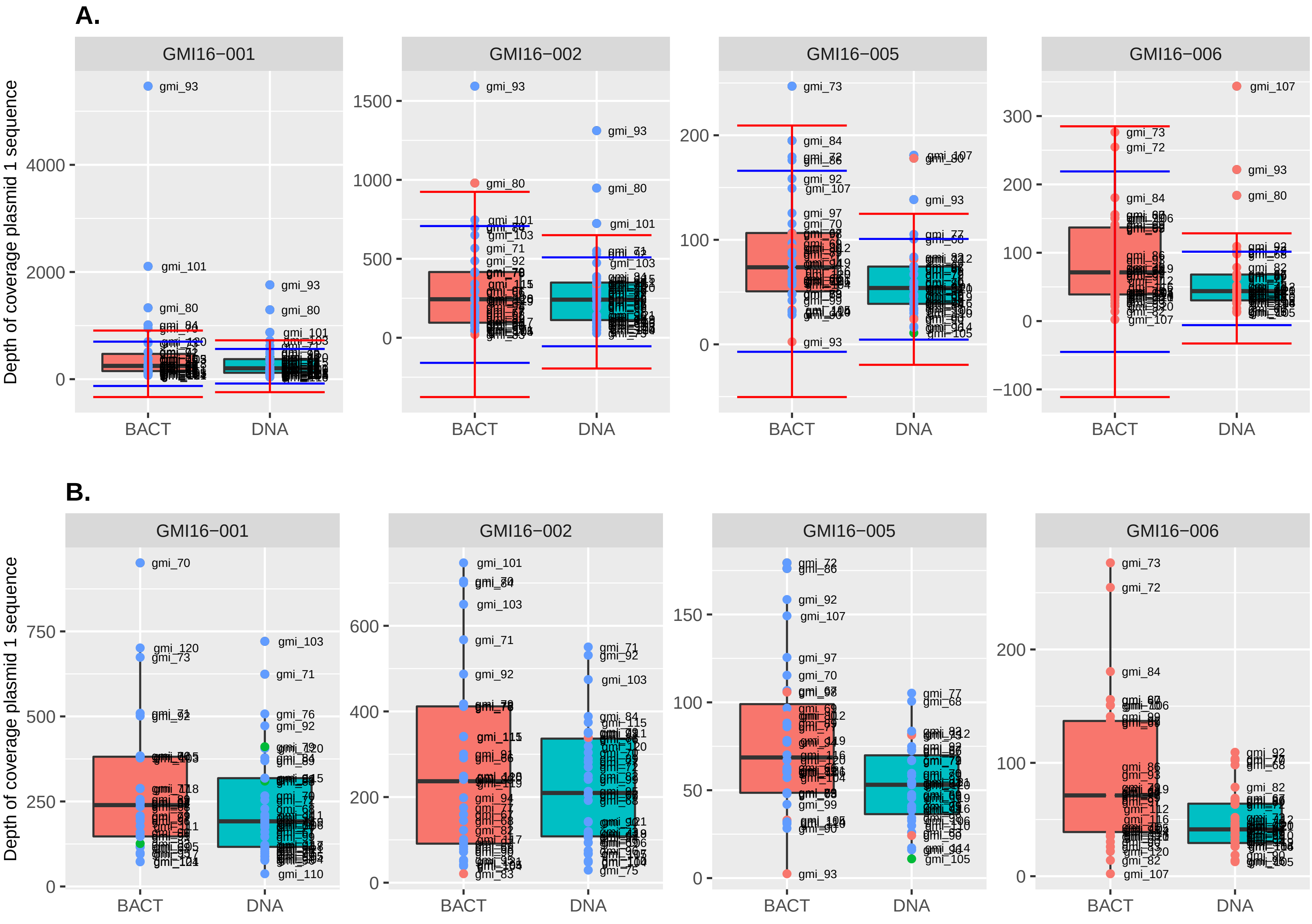
Depth of coverage total DNA sequence

GMI Proficiency Test 2016 Report - Figure 13



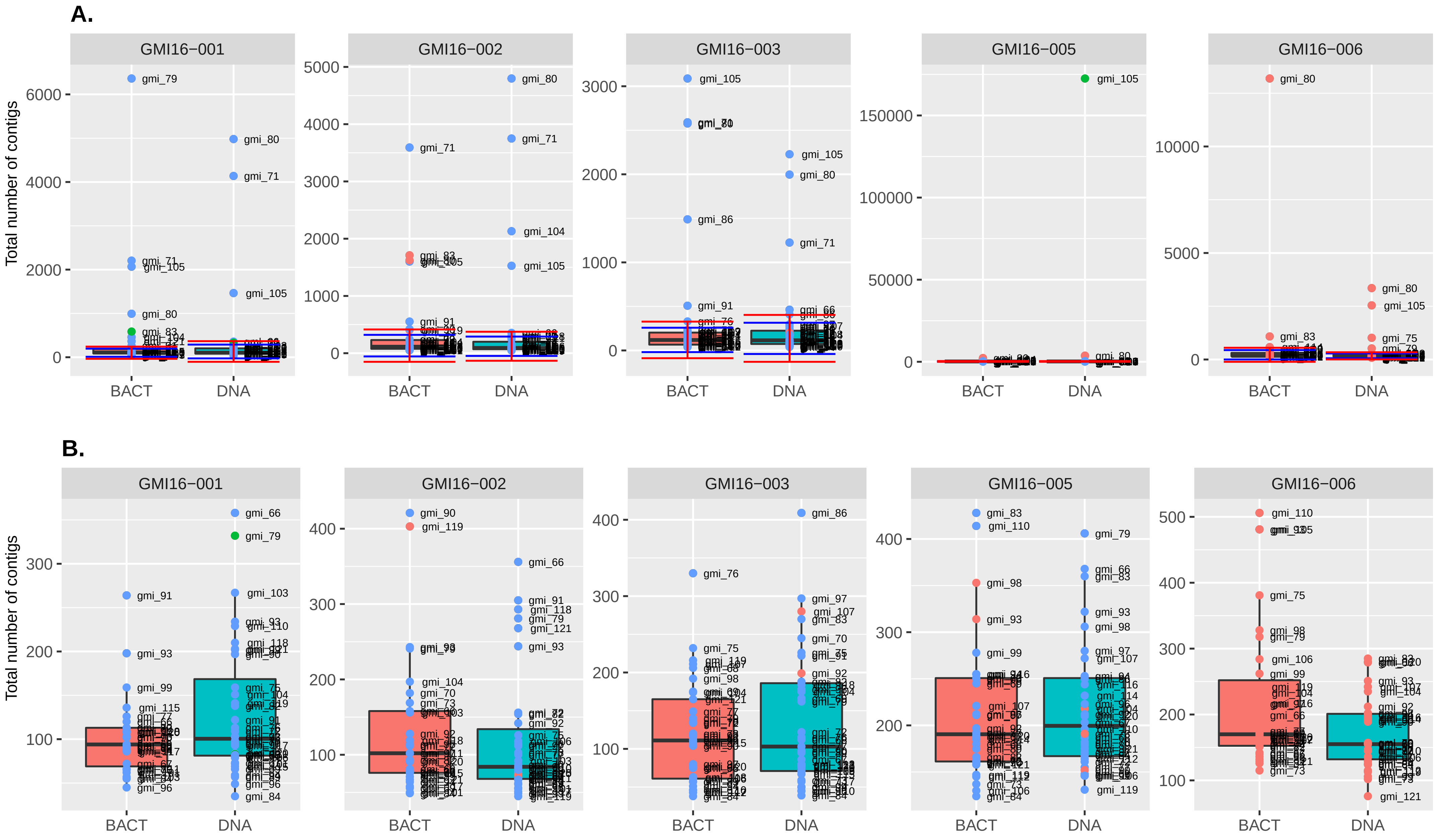
Depth of coverage plasmid 1 sequence

GMI Proficiency Test 2016 Report - Figure 14



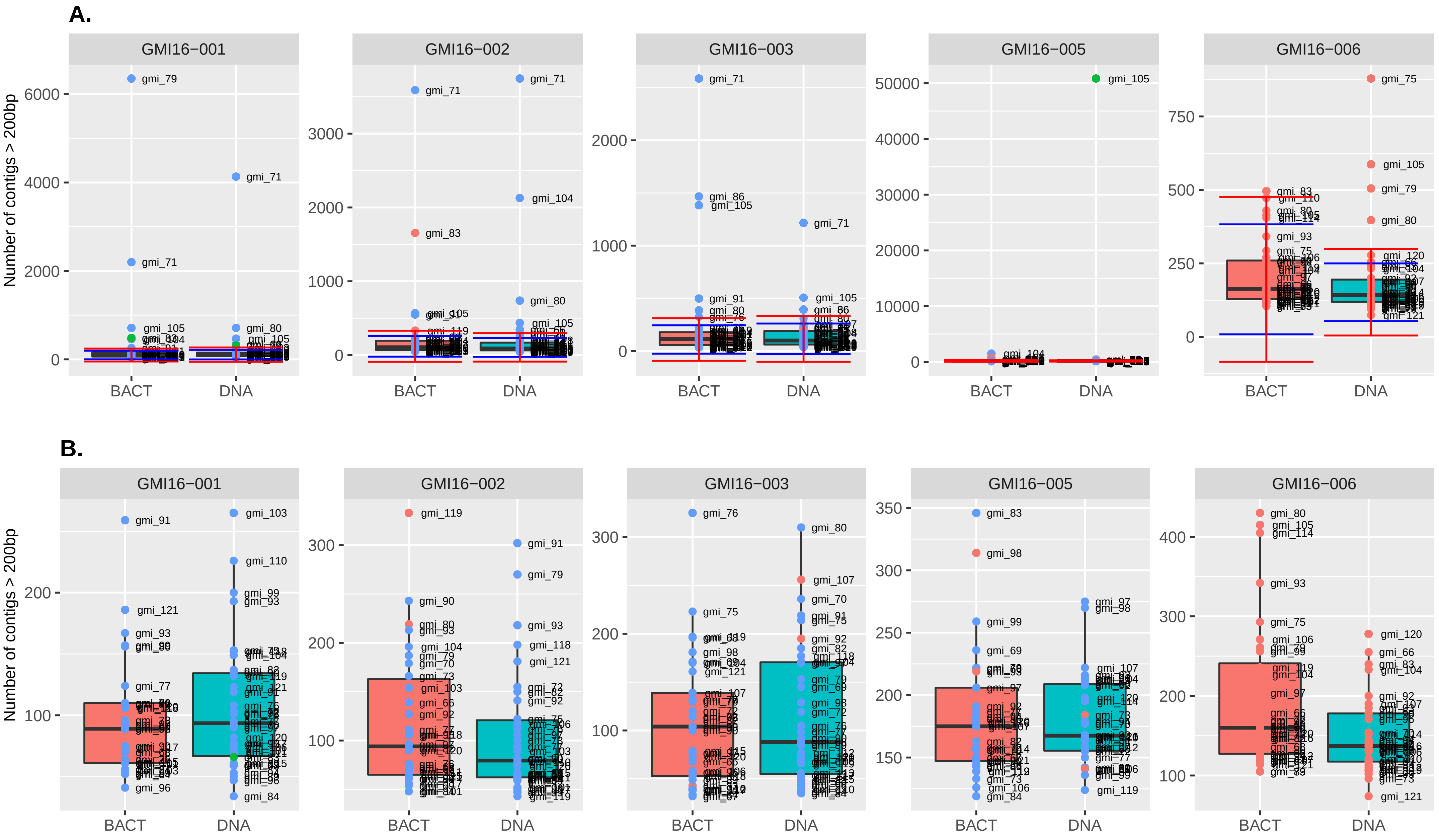
Total number of contigs

GMI Proficiency Test 2016 Report - Figure 15



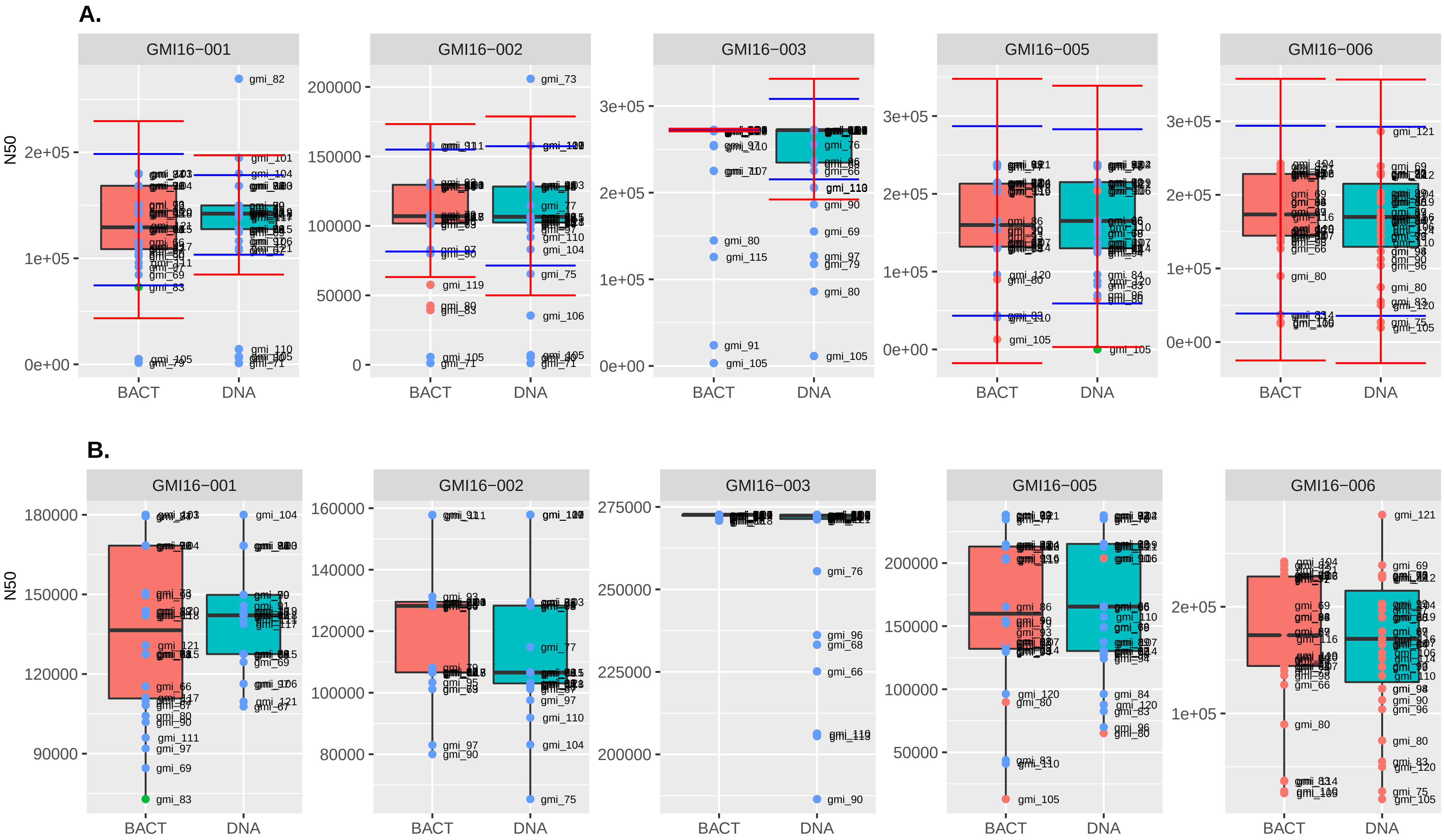
Number of contigs > 200bp

GMI Proficiency Test 2016 Report - Figure 16



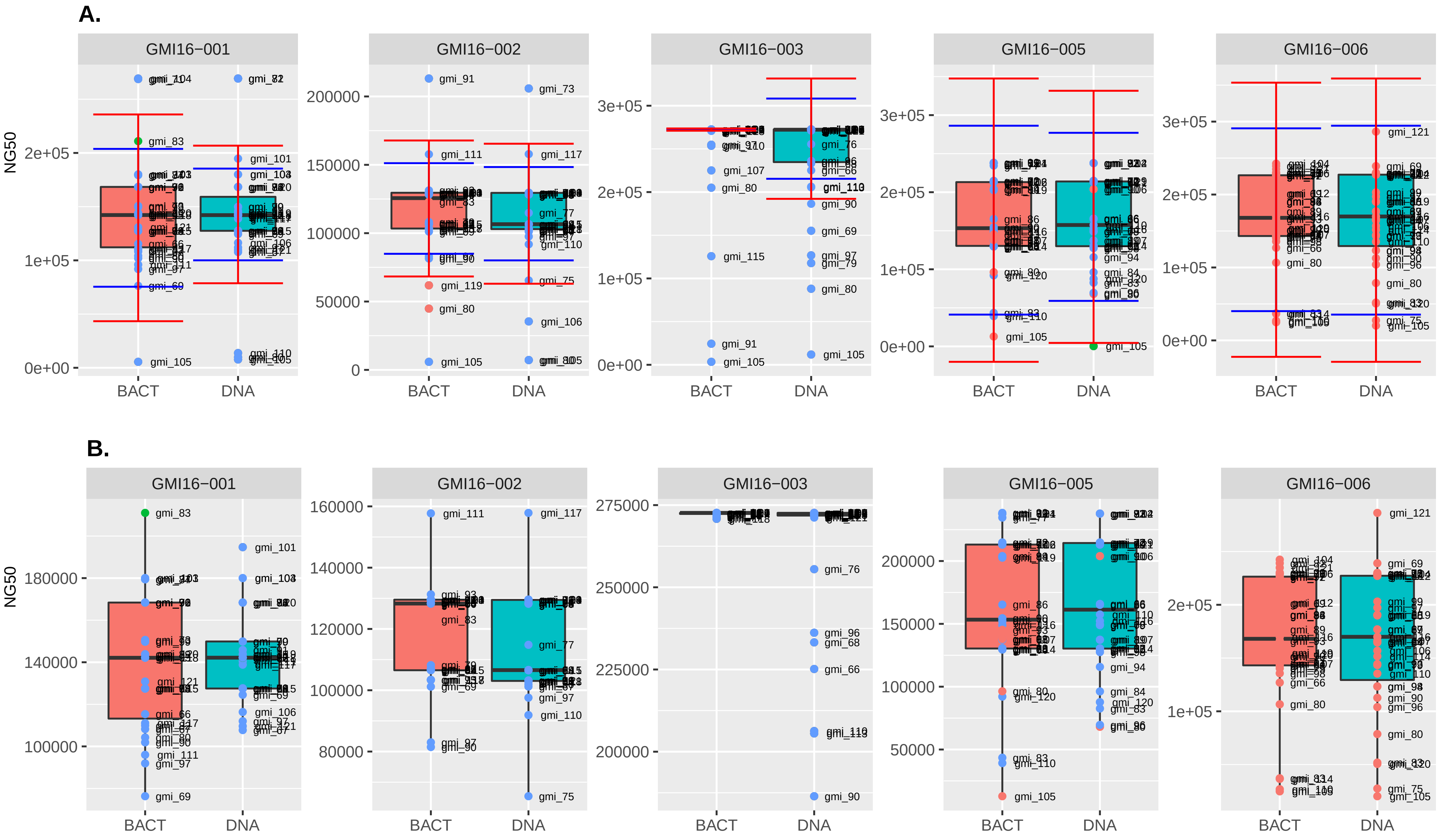
N50

GMI Proficiency Test 2016 Report - Figure 17

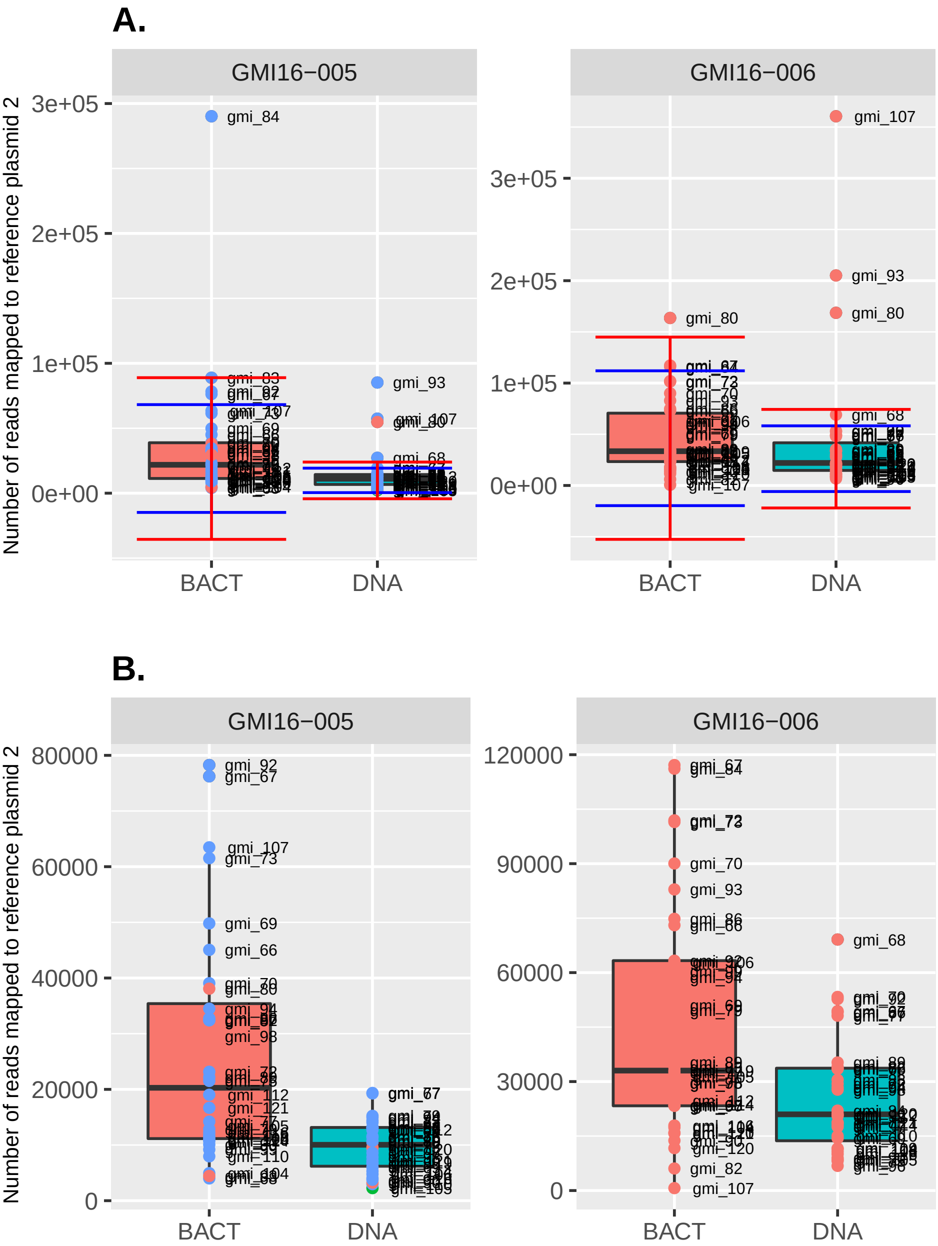


NG50

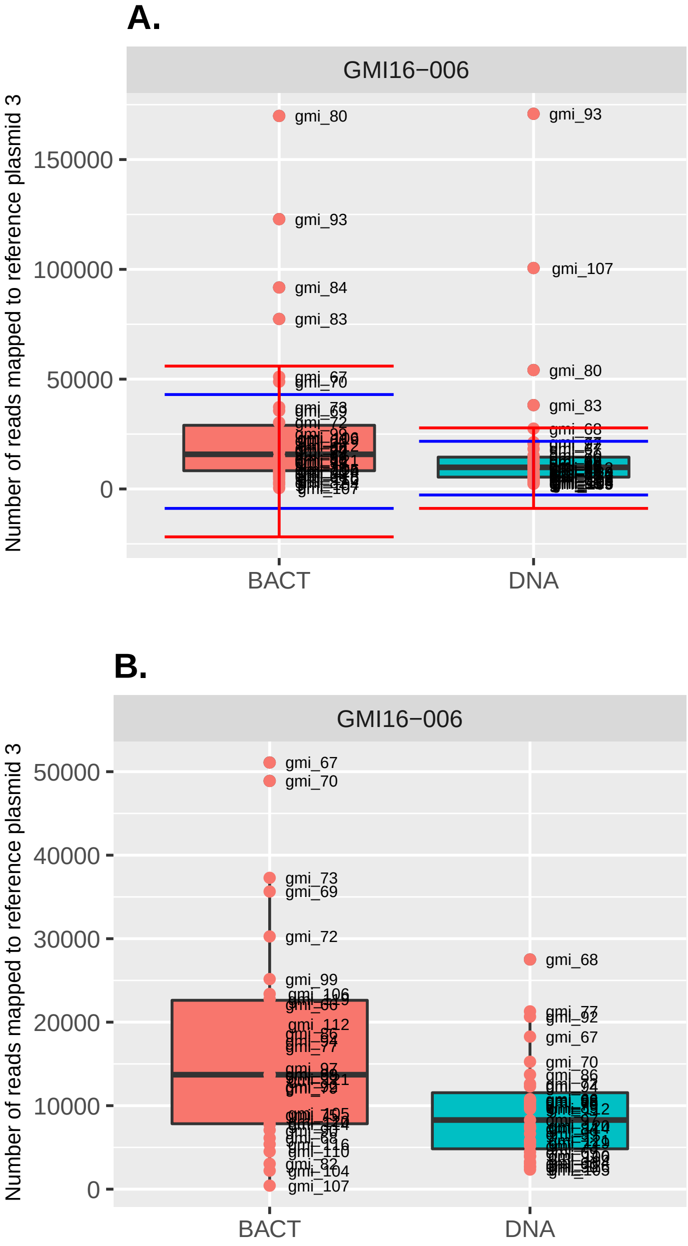
GMI Proficiency Test 2016 Report - Figure 18



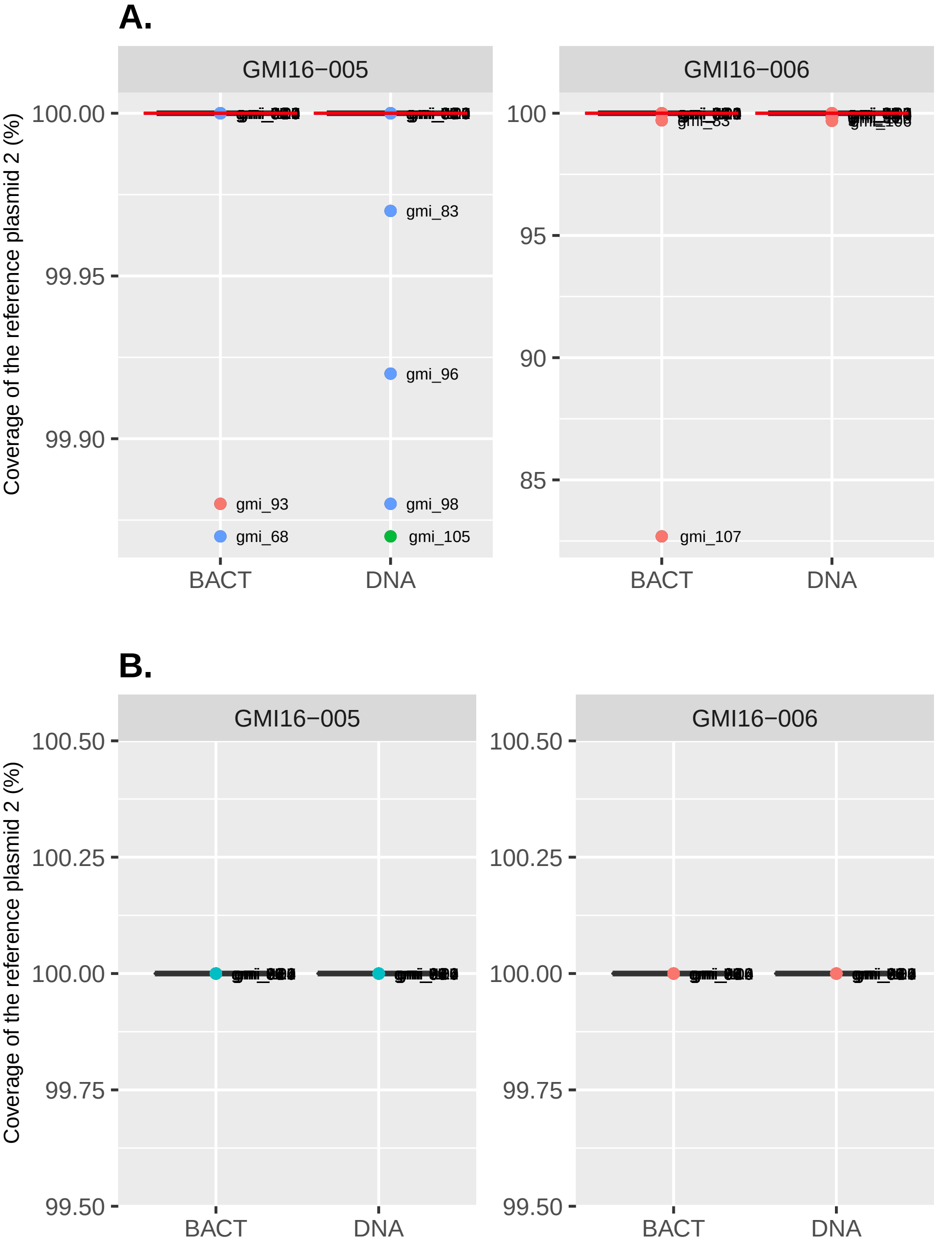
Number of reads mapped to reference plasmid 2



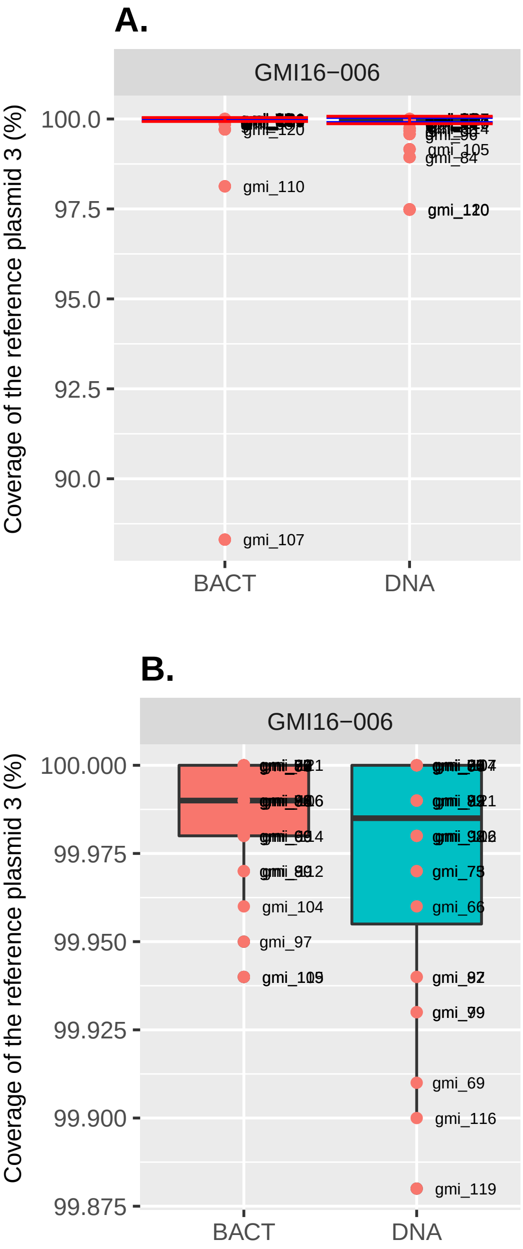
Number of reads mapped to reference plasmid 3



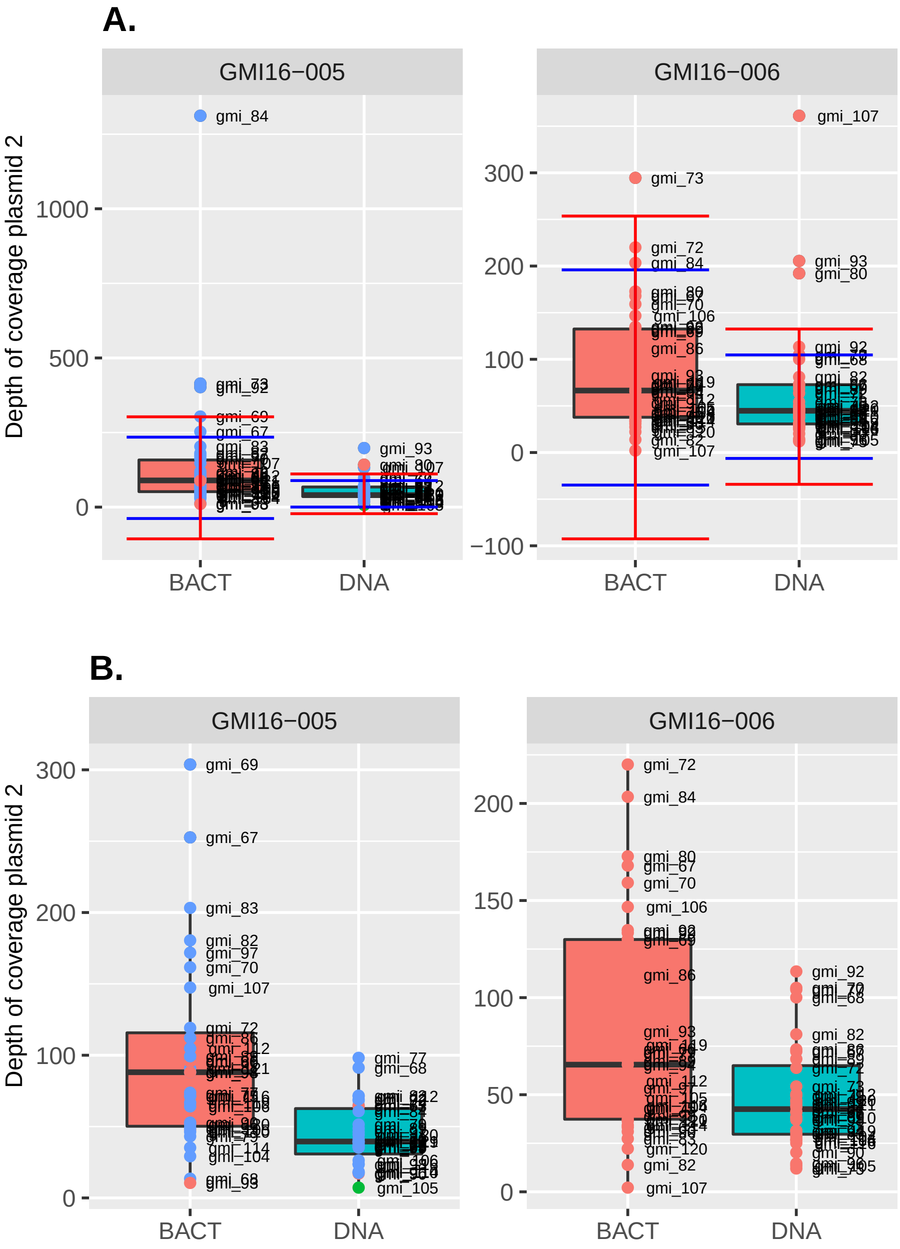
Coverage of the reference plasmid 2



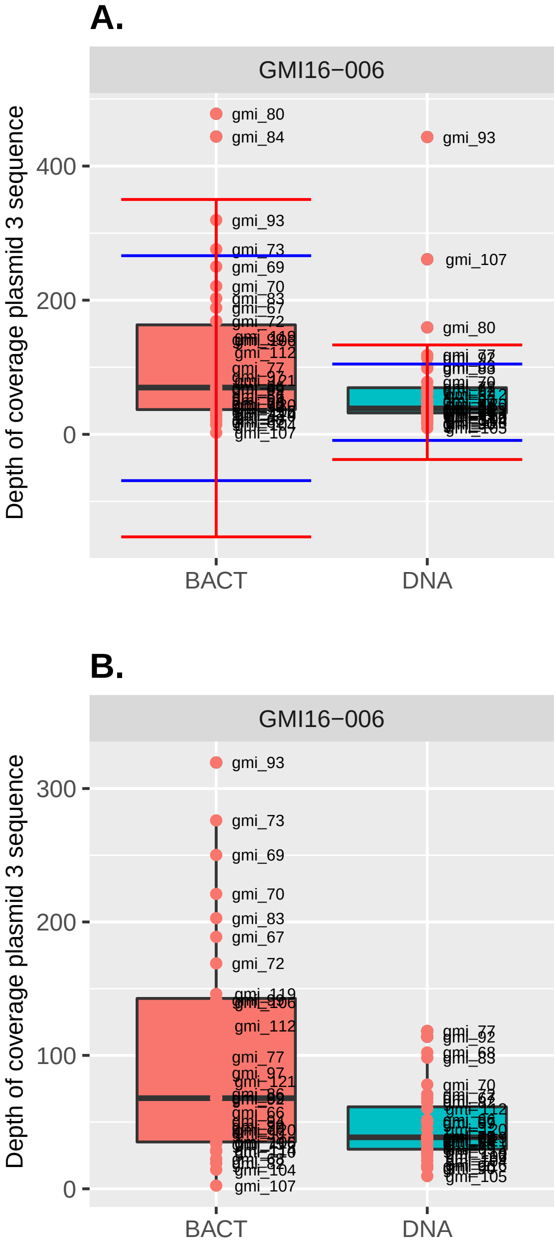
Coverage of the reference plasmid 3



Depth of coverage plasmid 2 sequence



Depth of coverage plasmid 3 sequence



Appendices

Appendix 1 Prenotification

Appendix 2 Protocol

Appendix 3 Letter to participants

Prenotification – GMI Proficiency Test 2016

GMI is a global, visionary taskforce of scientists and other stakeholders who share an aim of applying novel genomic technologies and informatics tools to improve 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 evaluation and production of reliable laboratory results of consistently good quality within the area of DNA preparation, sequencing, and analysis (e.g. clustering).

What is the GMI PT?

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

The GMI PT 2016 is 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 PT 2016 is supported by FDA GenomeTrakr and Microbiologics®.

The GMI working group 4 (WG4) steered by the US FDA, 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 identify variant sites within whole genome sequence data and cluster and distinguish samples based on those variants.

The proficiency test focuses on *Campylobacter coli* and *C. jejuni*, *Listeria monocytogenes* and *Klebsiella pneumoniae*, 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, participation in 1b is expected.

The three items consist of

> 1a) DNA extraction, purification, library-preparation, and whole-genome-sequencing of six bacterial cultures; one *Campylobacter coli* and one *C. jejuni* strain, two *Listeria monocytogenes* strains and two *Klebsiella pneumoniae* strains. Participants will be requested to submit reads using batch-upload via a web-interface and **optionally** also identify the Multi Locus Sequence Type (MLST) of the strains as well as the resistance genes present in the strains if the laboratory performs this type of analysis routinely.
>
> 1b) Perform whole-genome-sequencing of pre-prepared DNA delivered by GMI Working Group 4 of the same six bacterial strains mentioned in clause 1a.

2) Variant detection and phylogenetic/clustering analysis of three datasets each including fastq data from circa 20 genomes of *Campylobacter jejuni, Listeria monocytogenes and Klebsiella pneumoniae.* Note: If performing a reference based approach for variant detection, the reference applied for the analysis must be the species specific reference indicated in the proficiency test protocol.

For your information, this GMI PT 2016 is the second PT provided by the GMI initiative and will follow the same set-up as the one organized in 2015. Note, however, the GMI PT 2015 focused at *Salmonella enterica*, *Escherichia coli* and *Staphylococcus aureus.* The outcome of the GMI PT 2015 indicated that some of the applied quality markers could be species dependent. Therefore the GMI PT organizers selected the organisms *Campylobacter coli* and *C. jejuni*, *Listeria monocytogenes* and *Klebsiella pneumoniae* for the 2016 PT.

Who should participate?

All laboratories of the GMI community performing whole-genome-sequencing and/or phylogenetic/clustering analysis are invited to participate, in particular, laboratories frequently submitting data to NCBI, EBI and DDBJ.

Priority will be given to educational institutions and public health institutions. Private companies will not be accepted as participants. Should the sign up list exceed the number of participants that culture and DNA-material has been prepared for, participants will be included on a first come, first served basis.

Costs for participation?

There is no participation fee in the GMI proficiency test. The participating laboratories are, however, expected to cover all expenses as regards the handling and sequencing of the test strains and DNA in relation to their participation in the proficiency test. In addition, laboratories are expected to cover the expenses for parcel shipment if possible. Should a laboratory not be able to cover all expenses, they should notify the PT Coordinator in an email, as there might be limited funding available for this purpose.

The PT material will be sent as 'UN3373 Biological Substance Category B' (without temperature control). The courier selected to handle the bacterial cultures and DNA shipment must be able to handle UN3373 in your country, e.g. you could look into one of the following couriers which we have previously had good experience with: DHL, DHL Global Forwarding, Fedex, TNT or World Courier. We ask, also, that you provide your courier import account number in the sign-up form or directly to the PT Coordinator (please find contact information below). We need this information already at this stage to save time and resources.
Note: DHL distinguish between export account numbers and import account numbers; for the purpose of sending you the GMI PT bacterial cultures and DNA, we need your DHL import account number, i.e. the first digits must be 95 or 96, and the number must have 9 digits.

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 participate?
Via this link you access a sign-up webpage:
http://www.globalmicrobialidentifier.org/Workgroups/About-the-GMI-Proficiency-Test-2016

In the sign-up 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; *Campylobacter coli* and *C. jejuni*, *Listeria monocytogenes* and *Klebsiella pneumoniae*)

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

Protocol
The protocol including appendices is available for download.

Additional information relevant for participants will be sent directly by email or post.

Discussion forum
Note that a web-based discussion forum (https://foros.isciii.es/viewforum.php?f=7) is available for participants in the GMI PT 2016. The forum allows you to discuss issues relating to the analysis with other PT-participants. Appendix 4 of the protocol presents detailed information on the PT discussion forum.

Timeline
The bacterial isolates and the DNA will be shipped from DTU Food in October 2016. At this time, the data-files will also be available for download.

In order to minimize delays, **we ask you to 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 *Campylobacter* strains, two *Listeria monocytogenes* strains, two *Klebsiella pneumoniae* strains (note: these strains are carbapenemase-producing).

Deadline for submission of results
By **13 January 2017 (new extended deadline),** results must be submitted as described in the protocol. Individual results will be anonymized, and only the PT-organizers will have access to your laboratory's results. Each participating laboratory will receive an individual summary of the obtained performance. An overall report summarizing the results will be published and possibly subsequently in a peer-reviewed publication. Authors and co-authors of the publications will be those who have contributed to the preparation and execution of the proficiency test. Due to the anonymity of results, the individual participating laboratories will not be acknowledged in the publications.

Contact

If you have questions or comments to the GMI PT 2016, please contact the GMI PT Coordinator Susanne Karlsmose Pedersen (suska@food.dtu.dk)

PROTOCOL for GMI Proficiency Test, 2016

1 OVERVIEW AND OBJECTIVES

The proficiency test, 2016, 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**

NB: Please pay careful attention to instructions regarding the format and naming of results files submitted to the ftp site

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.

The GMI proficiency test 2016 is 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 PT 2016 is supported by the GenomeTrakr Network and Microbiologics®.

2 INTRODUCTION

GMI is a global, visionary taskforce of scientists and other stakeholders who share an aim of applying novel genomic technologies and informatics tools to improve 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.

The GMI working group 4 (WG4) steered by the US FDA, 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 identify variant sites within whole genome sequence data and cluster and distinguish samples based on those variants.

The proficiency test focuses on *Campylobacter jejuni*, *Listeria monocytogenes* and *Klebsiella pneumonia*, 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 also expected.

The three items consist of

1a) DNA extraction, purification, library-preparation, and whole-genome-sequencing of six bacterial cultures: two *Campylobacter* strains (one is *C. coli* and one is *C. jejuni*), two *Listeria monocytogenes* strains and two *Klebsiella pneumonia* 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 delivered by GMI Working Group 4 of the same six bacterial strains mentioned in clause 1a.

2) Variant detection and phylogenetic/clustering analysis of three datasets each including fastq data from circa 20 genomes of *C. jejuni*, *L. monocytogenes*, and *K. pneumonia*. Note: If performing a reference based approach for variant detection, the reference applied for the analysis must be the species specific references indicated below (see 3.4.2).

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 registered sign-up information.

3 OUTLINE OF THE GMI PT

3.1 Shipping, receipt and storage of bacterial strains

In October 2016, around 60 laboratories located worldwide will receive a parcel containing two strains of each *Campylobacter*, *L. monocytogenes*, and *K. pneumonia* together with corresponding purified DNA (according to the registered sign-up information). 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.

(c) if sequencing the samples is planned within the first 10 days of arrival of the shipment, you may store the dried samples in the zip-lock bag in which they arrived along with the silica gel desiccant pack. Should moisture start to appear, the desiccant pack must be changed.

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 KwikStik 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 (https://www.biomatrica.com/media/dnastable%20Plus/3001-0313.pdf). Each vial contains a minimum of 2 µg DNA.

Before use, the DNA should be rehydrated. Add 60 – 100 µl water or aqueous buffer to the dried DNA. Incubate the tubes at room temperature for 15 minutes, to allow complete hydration. Gently mix the sample by pipetting, to re-suspend the sample. The rehydrated DNA can be used directly in downstream application.

Unused rehydrated sample can be stored for up to one month at 4 degrees or room temperature.

Optional: The quality of the rehydrated DNA can be checked, by running it on a gel. The amount of DNA supplied in each tube will be sufficient to allow running a small fraction on a gel.

3.3.3 Item 2; Fastq data set

Three datasets, one for each of *C. jejuni* (cj), *L. monocytogenes* (lm), and *K. pneumonia* (kp) 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/r/PT_2016_bacterial_cultures_and_DNA (also see Appendix 2). Appendix 2 also describes the requested results when analyzing the sequences as regards the detected antimicrobial resistance genes and as regards the Multi Locus Sequence Type of the bacterial strain.

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 (cj, lm, or kp). Within each archive the participant will find the paired-end reads. The objective associated with this dataset is to assess the variability of laboratories in the clusters identified, which may be part of routine traceback investigations and/or source-tracking, 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).

If performing a reference based approach for variant detection, the reference applied for the analysis **must be** cj_reference.fasta (*C. jejuni*), lm_reference.fasta (*L. monocytogenes*) and kp_reference.fasta (*K. pneumonia*).

4 DISCUSSION FORUM

A web-based discussion forum (https://foros.isciii.es/viewforum.php?f=7) is available for participants in the GMI PT 2016, allowing for individual sign-up and discussion with other PT-participants and the PT organizers in relation to issues relating to the analysis for the present PT. Appendix 4 presents detailed information on the PT discussion forum. We strongly encourage the use of this forum as both a resource among participants but also as a platform to ask questions of the organizers – which other participants might also be interested in hearing the answer.

5 REPORTING OF RESULTS AND EVALUATION

For all items (1a, 1b and 2), the results should be captured and entered into the Internet-based survey (https://www.surveymonkey.com/r/PT_2016_bacterial_cultures_and_DNA and https://www.surveymonkey.com/r/PT_2016_FASTQ_dataset). See also Appendix 2 and 3.

5.1 Procedure and analysis of test material

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

Results for item 1a and 1b must be submitted as a batch-upload. The web-interface of the batch upload interface (https://cge.cbs.dtu.dk/services/ringtrials/) provides a possibility to upload several isolates in a single submission. The interface is divided into five steps, and a progress bar presents the overview of the submitted files.

Step 1; Login to the server using provided username (gmi_xx) and password (pink area).

Step 2; Download the Excel Metadata template to your computer (green area).

Step 3; Fill in the required fields with all the relevant information (metadata) about the isolates, the associated WGS file names, sequencing platform and sequencing type used to generate the data, etc. Each cell has a brief description of the required metadata. Some of the fields provide drop-down lists with possible metadata. Note that **sample name** should be the **same as label-name of the sample**, e.g. **GMI16-003-BACT** or **GMI16-003-DNA**.

Step 4; When the spreadsheet is properly filled out, upload the metadata file as well as the individual WGS files, that were included as metadata in the spreadsheet, to the web-interface by dropping the files to the ‘Drop metadata and sequence files here’ (grey area). After this, the spreadsheet will be validated to check if the metadata has the correct format.

If the spreadsheet contains invalid metadata, an error message will appear at the top of the uploader (between blue and green). To fix the errors, correct the errors in the original spreadsheet and upload the updated spreadsheet.

Step 5; Click on the Submit button to upload the files. The progress will be displayed in the Upload Progress bar (below the blue ‘Submit’ button). It is important to keep the window opened until the upload is completed. Pre-loading and validation will be displayed in blue and uploading – in green. After that, the web-interface will automatically submit the job to the server.

If the job is submitted correctly, you will get an on-screen confirmation message.

Via the Internet-based survey (https://www.surveymonkey.com/r/PT_2016_bacterial_cultures_and_DNA ; see also Appendix 2), answers should be submitted to the questions related to the analysed bacterial cultures and DNA.

5.1.2 Item 2; Fastq data set

For the fastq data set analyses, three types of files should be submitted and please pay close attention to the instructions on how samples and files should be named and which samples to include.

For each dataset:

1. The DNA sequence matrix used for clustering should be in fasta format and have that as the file extension
   - The matrix should only contain ***only*** those samples provided through the ftp site (i.e., there should be only 15 cj, 20 lm, and 17 kp samples in each file – no more no less). Please ensure that the fasta alignment has the exact number required by counting the number of sequence header prefixes (>).
   - Syntax for the names of samples in the matrix should be ***only*** the prefix preceding the first underscore in the file name. For example, **cj1_1.fastq** should be named **cj1** in the matrix.
   - The file should be named as follows **GMILabID_Taxon.fasta** (e.g., **GMI01_cj.FASTA**, **GMI01_lm.FASTA**, or **GMI01_kp.FASTA**).
2. The clusters themselves in newick format with the .tre as the file extension
   - Again, the tree should only contain those samples provided through the ftp site
   - Syntax for the names of samples in the tree file should be *only* the prefix preceding the first underscore in the file name. For example, **cj1_1.fastq** should be named **cj1** in the matrix.
   - The file should be named as follows **GMILabID_Taxon.tre** (e.g., **GMI01_cj.TRE**, **GMI01_lm.TRE**, or **GMI01_kp.TRE**).
   - Ideally the tree files would include branch lengths but this is not required.
3. The vcf (variant call format) files for each sample if a reference based approach was used and such files were produced.
   - The number of vcf files should match the number of samples found in the zipped archive from the ftp site.
   - Syntax for the names of the files should be *only* the prefix preceeding the first underscore in the file name. For example, **cj1_1.fastq** should be named **cj1** in the matrix.
   - The file should be named as follows **GMILabID_Taxon.tre** (e.g., **GMI01_cj1.vcf**, **GMI01_lm1.vcf**, or **GMI01_kp1.vcf**).

Via the Internet-based survey (https://www.surveymonkey.com/r/PT_2016_FASTQ_dataset; see also Appendix 3), answers should be submitted to the questions related to the Fastq data set section.

5.2 Evaluation of results

For both bacterial cultures and DNA (items 1a and 1b), the submitted sequence data (fastq-files) will be assembled using SPAdes [http://bioinf.spbau.ru/spades] and evaluated according to the following specific quality markers: e.g. 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.

Assessment of the submitted results from the analysis of the fastq datasets (item 2) is based on two criteria: 1) the concordance among laboratories in their answers to the questions in the SurveyMonkey (Appendix 3) and 2) the concordance between participants' in the information content contained in the SNP-matrix, vcf-files and the relationships among samples from the clustering analyses (i.e., the topology).

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.

5.3 Deadline for submission of results

Results must be submitted electronically **no later than December 14th, 2016**. Immediately after this date, the survey will be closed and results submitted to the Internet-based survey, via the batch-upload and to the ftp-site will be evaluated. Delayed submission of results will not be accepted.

5.4 Analysis and publication of results

Individual results will be anonymized, and only the PT-organizers will have access to your laboratory's results. Each participating laboratory will receive an individual summary of the obtained performance. An overall report summarizing the results will be published and subsequently in a peer-reviewed publication. Authors and co-authors of the publications will be those who have contributed to the preparation and execution of the proficiency test. Due to the anonymity of results, the individual participating laboratories will not be acknowledged in the publications.

We are looking forward to receiving your results.

If you have any questions or concerns, please do not hesitate to contact us, preferably by using the web-based discussion forum (https://foros.isciii.es/viewforum.php?f=7). In addition to receiving a response to your question, bringing up an issue via the forum allows other participants to also benefit from the discussions and the PT-organizers' response.

PT organizer related to the dry-lab fastq datasets:

James Pettengill
U.S. Food and Drug Administration
Center for Food Safety and Applied Nutrition
CPK1 RM2D0195100 Paint Branch Parkway
College Park, MD 20740, US
Tel: +1 240-402-1992
E-mail: James.Pettengill@fda.hhs.gov

PT organizer in relation to other issues, e.g. organizational issues, please contact the EQAS Coordinator:

Susanne Karlsmose Pedersen
National Food Institute, Technical University of Denmark
Søltofts Plads, Building 221, room 238,
DK-2800 Kgs. Lyngby, DENMARK
Tel: +45 3588 6601
E-mail: suska@food.dtu.dk

___ ___ ___

PROTOCOL for GMI Proficiency Test, 2016 - APPENDICES

Appendix 1: Using FTP to transfer files

Appendix 2: Overview of Internet-based survey - bacterial cultures and DNA

Appendix 3: Overview of Internet-based survey - FASTQ dataset

Appendix 4: GMI Proficiency Test Forum guide

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 2016 - bacterial cultures and DNA

Introduction

This survey seeks to capture info on participants' sequence procedures and specifications in relation to the bacterial cultures and DNA tested as part of the GMI Proficiency Test (PT) 2016.

The survey consists of six 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
6. Submitted datafiles

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

A web-based discussion forum (https://foros.isciii.es/viewforum.php?f=7) is also available for participants in the GMI PT 2016, allowing for individual sign-up and discussion with other PT-participants and the PT organizers in relation to issues relating to the analysis for the present PT. Appendix 4 of the PT protocol (see http://www.globalmicrobialidentifier.org/Workgroups/About-the-GMI-Proficiency-Test-2016) presents detailed information on the PT discussion forum. We strongly encourage the use of this forum as both a resource among participants but also as a platform to ask questions of the organizers – which other participants might also be interested in hearing the answer.

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

GMI is a global, visionary taskforce of scientists and other stakeholders who share an aim of applying novel genomic technologies and informatics tools to improve 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.

The GMI proficiency test 2016 is 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 PT 2016 is supported by the GenomeTrakr Network and Microbiologics®.

* 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
Date PT-material received		/		/	
Date processing the bacterial cultures started		/		/	
Date processing the bacterial cultures completed		/		/	
Date processing the DNA started		/		/	
Date processing the DNA completed (upload of sequence data)		/		/	

* 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

GMI Proficiency Testing 2016 - bacterial cultures and DNA

BACTERIAL CULTURES received

* 7. How were the bacterial cultures cultivated [as a decimal separator, please use full stop (.)]:

7.1 - Type of agar media/liquid broth:	
7.2 - Incubation time (hours):	
7.3 - Incubation temperature (˚C):	

* 8. For the Gram-negative bacterial cultures; DNA extraction procedure (enter 'NA' if not relevant):

8.1 - If manual extraction; kit used, full name:

8.2 - If manual extraction; catalogue number of kit:

8.3 - If manual extraction, modifications to kit protocol:

8.4 - If automatic extraction; robot used:

8.5 - If automatic extraction; specific protocol:

8.6 - If automatic extraction; modifications to protocol:

* 9. For the Gram-positive bacterial cultures; DNA extraction procedure (enter 'NA' if not relevant):

9.1 - If manual extraction; kit used, full name:

9.2 - If manual extraction; catalogue number of kit:

9.3 - If manual extraction, modifications to kit protocol:

9.4 - If automatic extraction; robot used:

9.5 - If automatic extraction; specific protocol:

9.6 - 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) [as a decimal separator, please use full stop (.)]

11.1 GMI16-001-BACT *(Campylobacter)*

11.2 GMI16-002-BACT *(Campylobacter)*

11.3 GMI16-003-BACT *(Listeria)*

11.4 GMI16-004-BACT *(Listeria)*

11.5 GMI16-005-BACT *(Klebsiella)*

11.6 GMI16-006-BACT *(Klebsiella)*

12. Total DNA amount (microgram) [as a decimal separator, please use full stop (.)]

12.1 GMI16-001-BACT (*Campylobacter*)

12.2 GMI16-002-BACT (*Campylobacter*)

12.3 GMI16-003-BACT *(Listeria)*

12.4 GMI16-004-BACT *(Listeria)*

12.5 GMI16-005-BACT *(Klebsiella)*

12.6 GMI16-006-BACT *(Klebsiella)*

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) [as a decimal separator, please use full stop (.)]

14.1 GMI16-001-BACT (*Campylobacter*)

14.2 GMI16-002-BACT (*Campylobacter*)

14.3 GMI16-003-BACT (*Listeria*)

14.4 GMI16-004-BACT (*Listeria*)

14.5 GMI16-005-BACT (*Klebsiella*)

14.6 GMI16-006-BACT (*Klebsiella*)

15. If relevant; measure of DNA quality (260/230 ratio) [as a decimal separator, please use full stop (.)]

15.1 GMI16-001-BACT (*Campylobacter*)

15.2 GMI16-002-BACT (*Campylobacter*)

15.3 GMI16-003-BACT (*Listeria*)

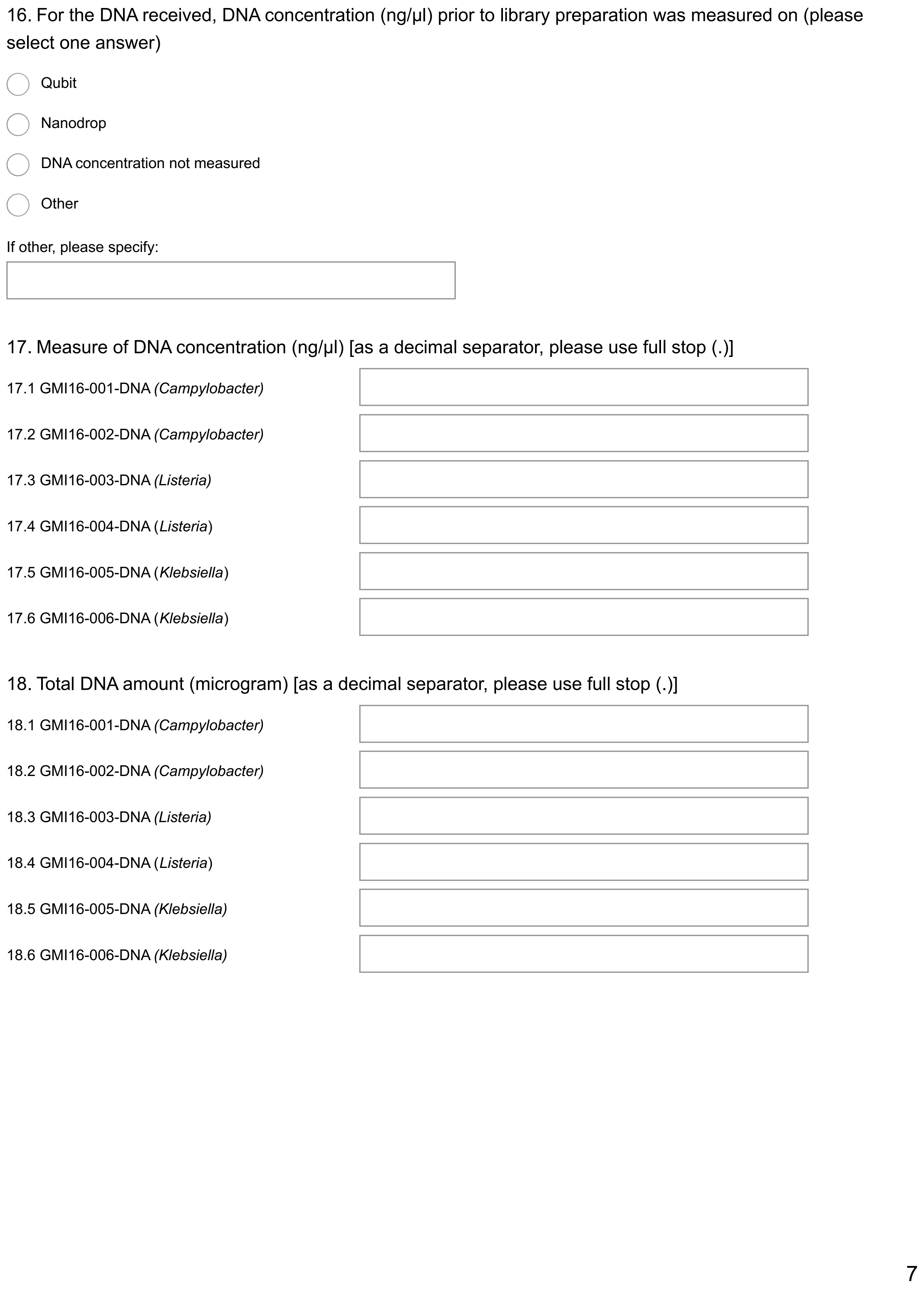
15.4 GMI16-004-BACT (*Listeria*)

15.5 GMI16-005-BACT (*Klebsiella*)

15.6 GMI16-006-BACT (*Klebsiella*)

GMI Proficiency Testing 2016 - bacterial cultures and DNA

DNA received



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) [as a decimal separator, please use full stop (.)]

17.1 GMI16-001-DNA *(Campylobacter)*	
17.2 GMI16-002-DNA *(Campylobacter)*	
17.3 GMI16-003-DNA *(Listeria)*	
17.4 GMI16-004-DNA (*Listeria*)	
17.5 GMI16-005-DNA (*Klebsiella*)	
17.6 GMI16-006-DNA (*Klebsiella*)	

18. Total DNA amount (microgram) [as a decimal separator, please use full stop (.)]

18.1 GMI16-001-DNA *(Campylobacter)*	
18.2 GMI16-002-DNA *(Campylobacter)*	
18.3 GMI16-003-DNA *(Listeria)*	
18.4 GMI16-004-DNA (*Listeria*)	
18.5 GMI16-005-DNA *(Klebsiella)*	
18.6 GMI16-006-DNA *(Klebsiella)*	

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:

20. Measure of DNA quality (e.g. RIN or 260/280 ratio) [as a decimal separator, please use full stop (.)]

20.1 GMI16-001-DNA (*Campylobacter*)

20.2 GMI16-002-DNA *(Campylobacter)*

20.3 GMI16-003-DNA *(Listeria)*

20.4 GMI16-004-DNA *(Listeria)*

20.5 GMI16-005-DNA *(Klebsiella)*

20.6 GMI16-006-DNA *(Klebsiella)*

21. If relevant; measure of DNA quality (260/230 ratio) [as a decimal separator, please use full stop (.)]

21.1 GMI16-001-DNA (*Campylobacter*)

21.2 GMI16-002-DNA *(Campylobacter)*

21.3 GMI16-003-DNA *(Listeria)*

21.4 GMI16-004-DNA *(Listeria)*

21.5 GMI16-005-DNA *(Klebsiella)*

21.6 GMI16-006-DNA *(Klebsiella)*

22. Did you perform quality check to verify the quality of the DNA on a gel (see description in the protocol of this optional check)

	Yes	No
22.1 GMI16-001-DNA (*Campylobacter*)	○	○
22.2 GMI16-002-DNA *(Campylobacter)*	○	○
22.3 GMI16-003-DNA *(Listeria)*	○	○
22.4 GMI16-004-DNA *(Listeria)*	○	○
22.5 GMI16-005-DNA *(Klebsiella)*	○	○
22.6 GMI16-006-DNA *(Klebsiella)*	○	○

GMI Proficiency Testing 2016 - bacterial cultures and DNA

SEQUENCING

23. 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; catalogue 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 □

* 24. 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 lix

○ MiSeq

○ MiSeq Dx

○ MiSeq FGx

○ ABI SOLiD

○ other

If other, please specify

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

○ Single-end

○ Paired-end

○ Not relevant

26. Sequencing details #2:
For the sequencing, the read length (bp) was set to be (expected read length)

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

[Note; this question refers to trimming performed actively by the participant (i.e. trimming performed automatically by your sequencer is not relevant for this question).
Ideally, no trimming should be performed.
As part of the analysis of the sequences subsequent to the deadline of the PT, trimming will be performed by application of the same tool for all submitted sequences. Should trimming be an integrated part of your sequencing process (disregarding possible automatic trimming by your sequencer), please indicate with 'yes' to this question]

○ Yes

○ No

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

28. For the analysis of the sequences from the bacterial cultures and the corresponding DNA 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:

GMI Proficiency Testing 2016 - bacterial cultures and DNA

ANALYSIS of sequences

29. If any, which method was used to characterize or differentiate isolates (please select all that apply)?

□ MLST

□ Allele-based

□ Gene-by-gene-based

□ SNP-based

□ None

Other (please specify)

30. If you determined the MLST-type of the sequenced DNA, how was the analysis performed (please select one answer)?

○ MLST-analysis was performed on raw reads

○ MLST-analysis was performed on contigs

○ MLST-analysis was not performed

31. If you determined antimicrobial resistance (AMR) genes present in the sequenced DNA, how was the analysis performed (please select one answer)?

○ Analysis for AMR-genes was performed on raw reads

○ Analysis for AMR-genes was performed on contigs

○ Analysis for AMR-genes was not performed

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

32.1 GMI16-001-BACT (*Campylobacter*)

32.2 GMI16-002-BACT *(Campylobacter)*

32.3 GMI16-003-BACT (*Listeria*)

32.4 GMI16-004-BACT (*Listeria*)

32.5 GMI16-005-BACT (*Klebsiella*)

32.6 GMI16-006-BACT (*Klebsiella*)

33. For the DNA from the received bacterial culture, if MLST-analysis was performed based on the sequence analysis, which alleles characterize the isolate?

33.1 GMI16-001-BACT (*Campylobacter*)

33.2 GMI16-002-BACT *(Campylobacter)*

33.3 GMI16-003-BACT (*Listeria*)

33.4 GMI16-004-BACT (*Listeria*)

33.5 GMI16-005-BACT *(Klebsiella)*

33.6 GMI16-006-BACT (*Klebsiella*)

34. 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 (please list the genes according to the following order of antimicrobial classes: Aminocyclitols, aminoglycosides, ß-lactams, fluoroquinolones, glycopeptides, lincosamides, macrolides, oxazolidones, phenicols, pleuromutilins, polypeptide antibiotics, quinolones, streptogramins, sulfonamides, tetracyclines, trimethoprim, other)?

34.1 GMI16-001-BACT *(Campylobacter)*

34.2 GMI16-002-BACT *(Campylobacter)*

34.3 GMI16-003-BACT *(Listeria)*

34.4 GMI16-004-BACT (*Listeria*)

34.5 GMI16-005-BACT *(Klebsiella)*

34.6 GMI16-006-BACT *(Klebsiella)*

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

35.1 GMI16-001-DNA (*Campylobacter*)

35.2 GMI16-002-DNA (*Campylobacter*)

35.3 GMI16-003-DNA (*Listeria*)

35.4 GMI16-004-DNA (*Listeria*)

35.5 GMI16-005-DNA (*Klebsiella*)

35.6 GMI16-006-DNA (*Klebsiella*)

36. For the received DNA, if MLST-analysis was performed based on the sequence analysis, which alleles characterize the isolate?

36.1 GMI16-001-DNA (*Campylobacter*)

36.2 GMI16-002-DNA *(Campylobacter)*

36.3 GMI16-003-DNA *(Listeria)*

36.4 GMI16-004-DNA *(Listeria)*

36.5 GMI16-005-DNA *(Klebsiella)*

36.6 GMI16-006-DNA *(Klebsiella)*

37. 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 (please list the genes according to the following order of antimicrobial classes: Aminocyclitols, aminoglycosides, ß-lactams, fluoroquinolones, glycopeptides, lincosamides, macrolides, oxazolidones, phenicols, pleuromutilins, polypeptide antibiotics, quinolones, streptogramins, sulfonamides, tetracyclines, trimethoprim, other)?

37.1 GMI16-001-DNA (*Campylobacter*)

37.2 GMI16-002-DNA (*Campylobacter*)

37.3 GMI16-003-DNA *(Listeria)*

37.4 GMI16-004-DNA (*Listeria*)

37.5 GMI16-005-DNA (*Klebsiella*)

37.6 GMI16-006-DNA (*Klebsiella*)

38. 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:

39. 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 2016 - bacterial cultures and DNA

SUBMITTED datafiles

* 40. The obtained non-assembled sequence data have been uploaded for bacterial cultures and DNA following the description of batch-upload in the PT-protocol, for

	Yes	No
Campylobacter	○	○
Listeria	○	○
Klebsiella	○	○

Comments:

□

Info:
For both bacterial cultures and DNA, the submitted sequence data (fastq-files) will be evaluated according to the following specific quality markers, e.g. 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 for both the bacterial cultures and DNA.

GMI Proficiency Testing 2016 - FASTQ dataset

Introduction

This survey seeks to capture info in relation to the fastq data set component of the GMI Proficiency test (PT) 2016.

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

A web-based discussion forum (https://foros.isciii.es/viewforum.php?f=7) is also available for participants in the GMI PT 2016, allowing for individual sign-up and discussion with other PT-participants and the PT organizers in relation to issues relating to the analysis for the present PT. Appendix 4 of the PT protocol (see http://www.globalmicrobialidentifier.org/Workgroups/About-the-GMI-Proficiency-Test-2016) presents detailed information on the PT discussion forum. We strongly encourage the use of this forum as both a resource among participants but also as a platform to ask questions of the organizers – which other participants might also be interested in hearing the answer.

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

GMI is a global, visionary taskforce of scientists and other stakeholders who share an aim of applying novel genomic technologies and informatics tools to improve 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.

The GMI proficiency test 2016 is 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 PT 2016 is supported by the GenomeTrakr Network and Microbiologics®.

* 1. Institute name / Organization name

* 2. Department name

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

GMI Proficiency Testing 2016 - FASTQ dataset

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 variant detection, which of the following did you use:
(please select one answer)

○ A reference based approach

○ De novo assemblies

○ A combination of both

* 7. If you perform *de novo* assemblies, which tool did you apply (please insert name and URL; e.g. Velvet, https://www.ebi.ac.uk/~zerbino/velvet/)?

(Enter 'NA' if you do not do *de novo* assemblies)

* 8. If you use a reference-based approach, which tools do you use for mapping and variant detection (please insert name and URL; e.g. Bowtie2, http://bowtie-bio.sourceforge.net/bowtie2/index.shtml; VarScan, http://varscan.sourceforge.net)?

(Enter 'NA' if you do not use a reference-based approach)

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

○ SNPs

○ Methodology other than SNPs

If methodology other than SNPs (please specify):

If applying SNPs, go to question 10,
If not applying SNPs, go to question 12

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

10.1 - *Campylobacter*

10.2 - *Listeria*

10.3 - *Klebsiella*

11. Which criteria did you use for SNPs filtering:

11.1 - Filter SNPs with excess coverage (i.e. repetitive regions):

11.2 - Did you filter SNPs occurring in a cluster (a.k.a. pruning) (indicate 'yes' or 'no'):

11.3 - Which definition of the cluster did you use (i.e. ≥3 SNPs in 1000 base pairs (bp):

11.4 - Other, please specify:

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

13. 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 (see description in the PT-protocol)

14. If you do assemblies, do you calculate the number of contigs (please select one answer)

○ Yes

○ No

○ We don't perform assemblies

15. If you do assemblies, do you filter out contigs below a certain size (please select one answer)

○ Yes

○ No

○ We don't perform assemblies

If yes, indicate minimum size

16. If you do assemblies, do you calculate N50 (please select one answer)

○ Yes

○ No

○ We don't perform assemblies

17. If you perform assemblies, do you calculate the size of the chromosome (please select one answer)

○ Yes

○ No

○ We don't perform assemblies

18. Do you calculate coverage as a quality metric? (please select one answer)

○ Yes

○ No

* 19. Did you check for contamination and/or verify the species? If so, 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)

(Enter 'NA' if you do not check for contamination and/or verify the species)

20. If you did check, could you verify the species?

□ We did not attempt to verify species

□ Yes, for all *Campylobacter*

□ Yes, for all *Listeria*

□ Yes, for all *Klebsiella*

□ No, for some *Campylobacter*

□ No, for some *Listeria*

□ No, for some *Klebsiella*

If no, please indicate why

□

GMI Proficiency Testing 2016 - FASTQ dataset

SUBMITTED datafiles

Please carefully follow the instructions regarding the naming of submitted files and the samples that should be included in them! Thank you.

* 21. The following files have been submitted to the ftp-site:

	Campylobacter dataset	*Listeria* dataset	*Klebsiella* dataset
A fasta formatted DNA sequence matrix that was used for clustering (e.g., a fasta file extension)	⬍	⬍	⬍
A newick formatted file with the clusters themselves (.tre file extension) (the format can be obtained through the R package APE or using FigTree's "Export Trees" option)	⬍	⬍	⬍
vcf (variant call format) files for each sample if a reference based approach was used and such files were produced	⬍	⬍	⬍

Comments:

□

The number and identity of samples in each uploaded file should match exactly those that were included in the original data (i.e., 15, 20, 17 sequences of *Campylobacter jejuni*, *Listeria monocytogenes* and *Klebsiella pneumoniae*, respectively). Within each file the samples should be named as the prefix within the original fastq file and the file should be named GMILabID_Taxon.appropriateFileExtenation (e.g., GMI01_cj.fasta). See the protocol for specific instructions and additional information on the samples to be included and the naming of them.

GMI Proficiency Test Forum guide

Global Microbial Identifier

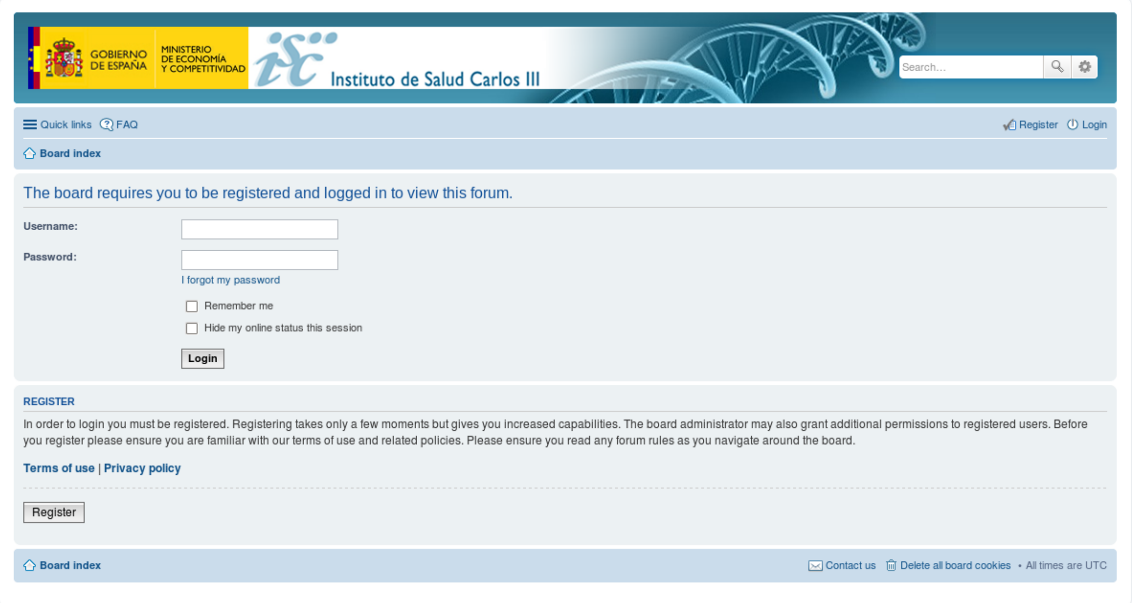
Content

The purpose of the *GMI Profiency Test* forum is to facilitate the exchange of information and experiences among GMI PT participants.

Sign up to forum platform

GMI Profiency Test forum is hosted by the *Instituto de Salud Carlos III* (*Spanish National Health Institute Carlos III*). The first step is to register and create an account, which is described below.

Go to the GMI PT forum web page (*https://foros.isciii.es/viewforum.php?f=7*) and press the ***Register*** button.


GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
Instituto de Salud Carlos III
Search…
Quick links
FAQ
Register
Login
Board index

The board requires you to be registered and logged in to view this forum.

Username:

Password:

I forgot my password

□ Remember me

□ Hide my online status this session

Login

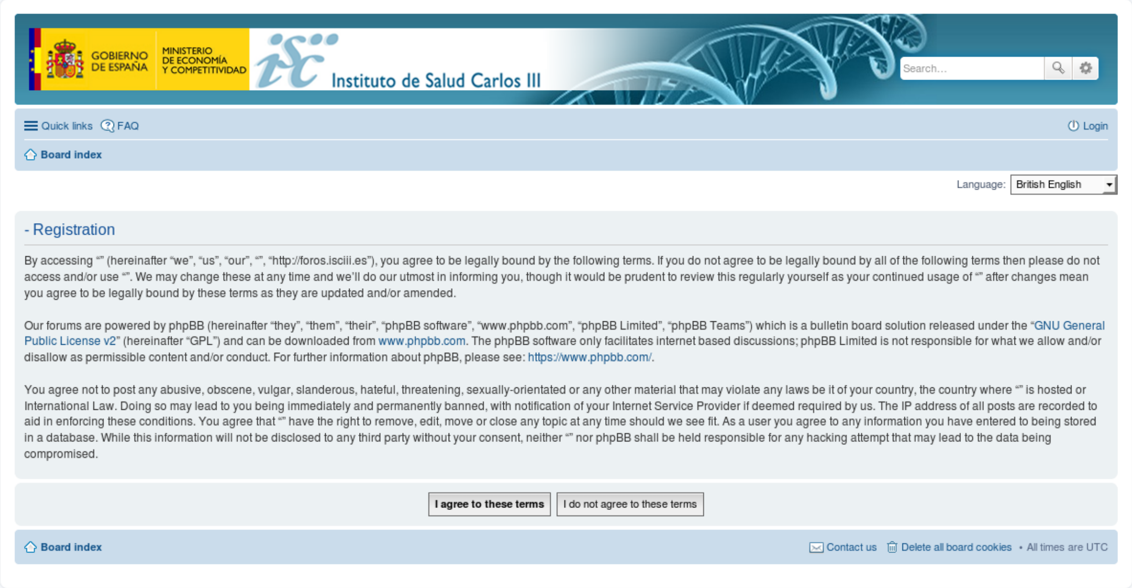
REGISTER

In order to login you must be registered. Registering takes only a few moments but gives you increased capabilities. The board administrator may also grant additional permissions to registered users. Before you register please ensure you are familiar with our terms of use and related policies. Please ensure you read any forum rules as you navigate around the board.

Terms of use | Privacy policy

Register

Read carefully forum terms and press *I agree to these terms* to accept agreement.


GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
Instituto de Salud Carlos III
Search…
Quick links
FAQ
Login
Board index

Language: British English

- Registration

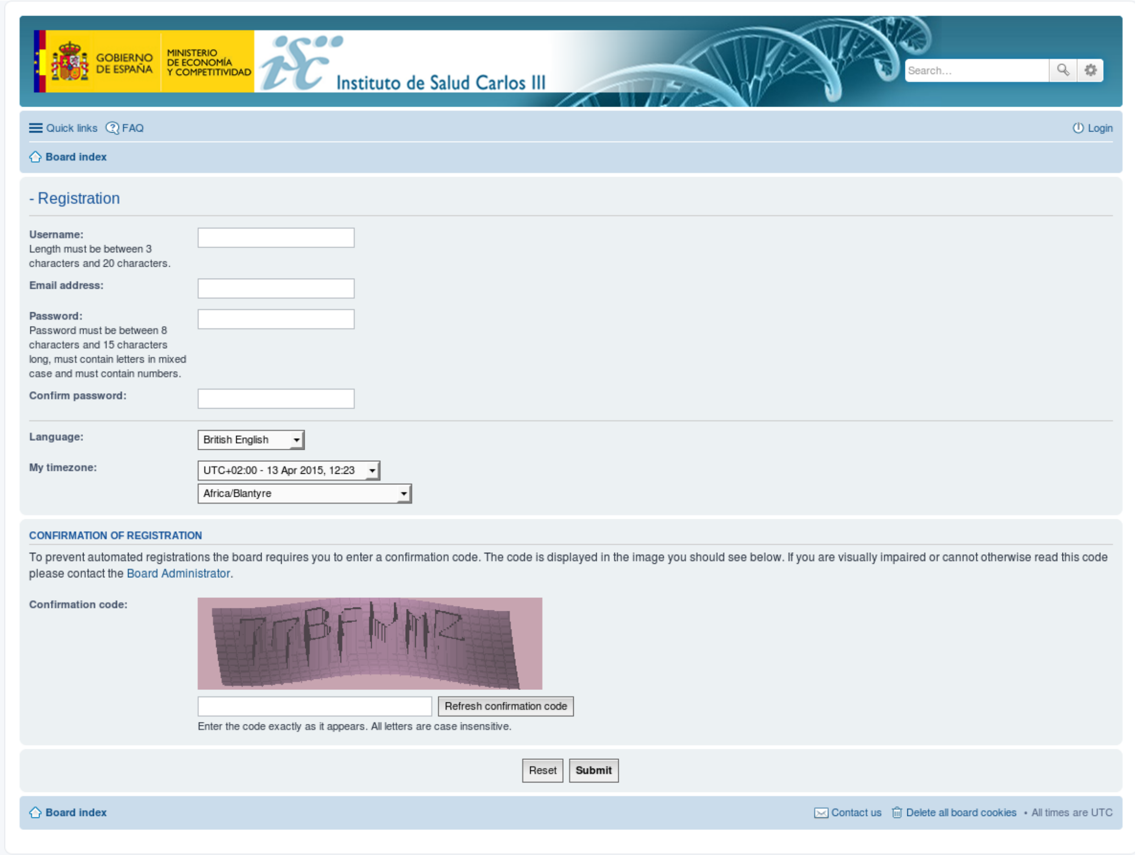
By accessing “” (hereinafter “we”, “us”, “our”, “”, “http://foros.isciii.es”), you agree to be legally bound by the following terms. If you do not agree to be legally bound by all of the following terms then please do not access and/or use “”. We may change these at any time and we'll do our utmost in informing you, though it would be prudent to review this regularly yourself as your continued usage of “” after changes mean you agree to be legally bound by these terms as they are updated and/or amended.

Our forums are powered by phpBB (hereinafter “they”, “them”, “their”, “phpBB software”, “www.phpbb.com”, “phpBB Limited”, “phpBB Teams”) which is a bulletin board solution released under the “GNU General Public License v2” (hereinafter “GPL”) and can be downloaded from www.phpbb.com. The phpBB software only facilitates internet based discussions; phpBB Limited is not responsible for what we allow and/or disallow as permissible content and/or conduct. For further information about phpBB, please see: https://www.phpbb.com/.

You agree not to post any abusive, obscene, vulgar, slanderous, hateful, threatening, sexually-orientated or any other material that may violate any laws be it of your country, the country where “” is hosted or International Law. Doing so may lead to you being immediately and permanently banned, with notification of your Internet Service Provider if deemed required by us. The IP address of all posts are recorded to aid in enforcing these conditions. You agree that “” have the right to remove, edit, move or close any topic at any time should we see fit. As a user you agree to any information you have entered to being stored in a database. While this information will not be disclosed to any third party without your consent, neither “” nor phpBB shall be held responsible for any hacking attempt that may lead to the data being compromised.

I agree to these terms
I do not agree to these terms
Board index
Contact us
Delete all board cookies • All times are UTC

Enter your username (with a length between 3 and 20 characters), email address and set up a password (must be between 8 and 15 characters long, must contain letters in mixed case and numbers). Additionally you could change your time zone. Enter confirmation code and press ***Submit***.


GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
Instituto de Salud Carlos III
Search…
Quick links
FAQ
Login
Board index

- Registration

Username:
Length must be between 3 characters and 20 characters.

Email address:

Password:
Password must be between 8 characters and 15 characters long, must contain letters in mixed case and must contain numbers.

Confirm password:

Language: British English

My timezone: UTC+02:00 - 13 Apr 2015, 12:23
Africa/Blantyre

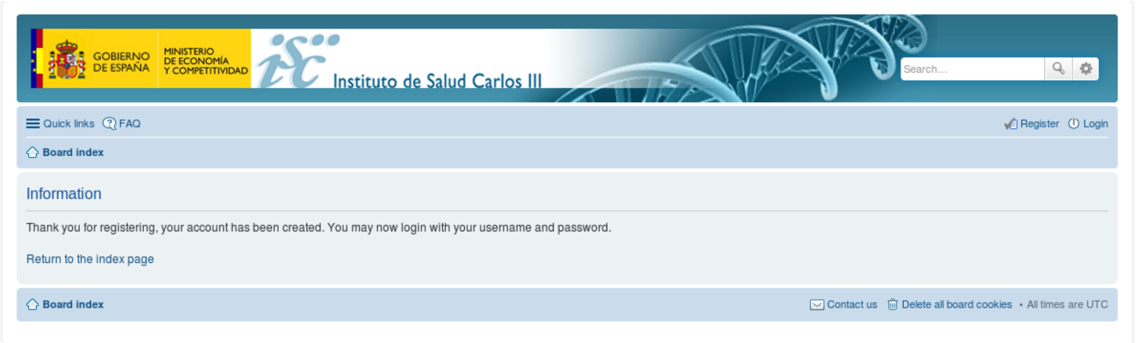
CONFIRMATION OF REGISTRATION

To prevent automated registrations the board requires you to enter a confirmation code. The code is displayed in the image you should see below. If you are visually impaired or cannot otherwise read this code please contact the Board Administrator.

Confirmation code:

Enter the code exactly as it appears. All letters are case insensitive.

A welcome email will be sent to your email address. Click on ***Return to the index page*** and log in using your selected username and password.


GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
Instituto de Salud Carlos III
Search…
Quick links
FAQ
Register
Login
Board index

Information

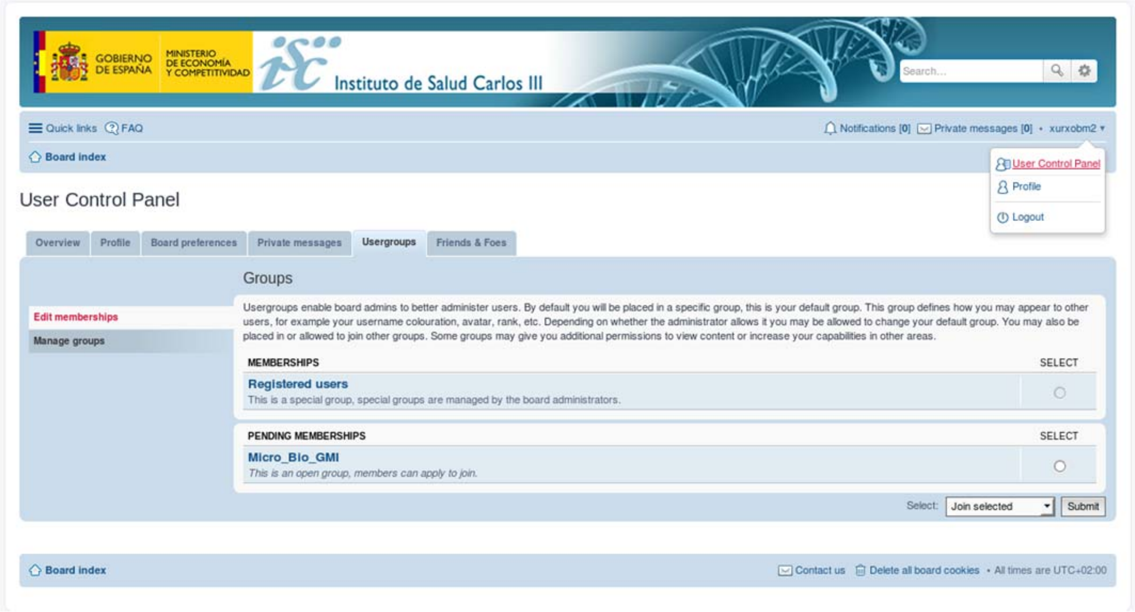
Thank you for registering, your account has been created. You may now login with your username and password.

Return to the index page

Request to join GMI PT forum group

Joining as a member of GMI PT forum requires to be accepted by *Micro_Bio_GMI* group. Membership is restricted to *GMI Proficiency Test* participants. This section details the procedure to request that you be added to that group.

Click on a link located at the upper right corner of the screen that is labeled with your username. Click on ***User Control Panel*** and select ***Usergroups*** tab.



User Control Panel

Groups

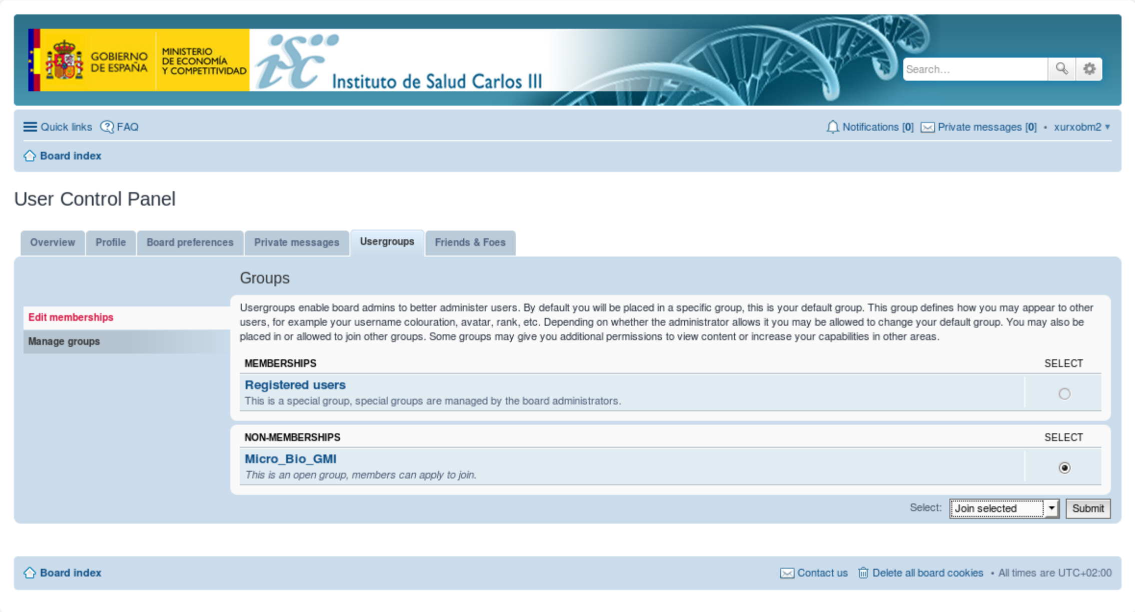
Usergroups enable board admins to better administer users. By default you will be placed in a specific group, this is your default group. This group defines how you may appear to other users, for example your username colouration, avatar, rank, etc. Depending on whether the administrator allows it you may be allowed to change your default group. You may also be placed in or allowed to join other groups. Some groups may give you additional permissions to view content or increase your capabilities in other areas.

MEMBERSHIPS	SELECT
Registered users This is a special group, special groups are managed by the board administrators.	○

PENDING MEMBERSHIPS	SELECT
Micro_Bio_GMI *This is an open group, members can apply to join.*	○

Select: Join selected ▾ Submit

Clik on ***Micro_Bio_GMI*** radio button. Select ***Join selected*** in *Select* dropdown list and press ***Submit*** button.



User Control Panel

Groups

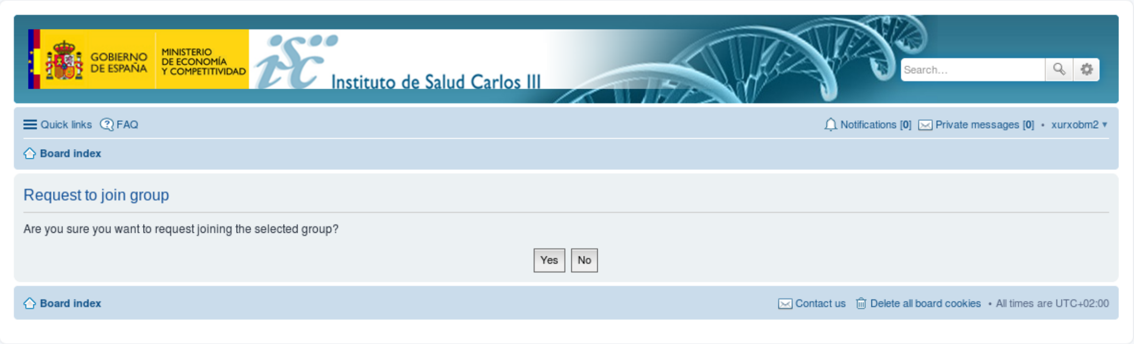
Usergroups enable board admins to better administer users. By default you will be placed in a specific group, this is your default group. This group defines how you may appear to other users, for example your username colouration, avatar, rank, etc. Depending on whether the administrator allows it you may be allowed to change your default group. You may also be placed in or allowed to join other groups. Some groups may give you additional permissions to view content or increase your capabilities in other areas.

MEMBERSHIPS	SELECT
Registered users This is a special group, special groups are managed by the board administrators.	○

NON-MEMBERSHIPS	SELECT
Micro_Bio_GMI *This is an open group, members can apply to join.*	●

Select: Join selected ▾ Submit

Confirm request to GMI PT forum group.



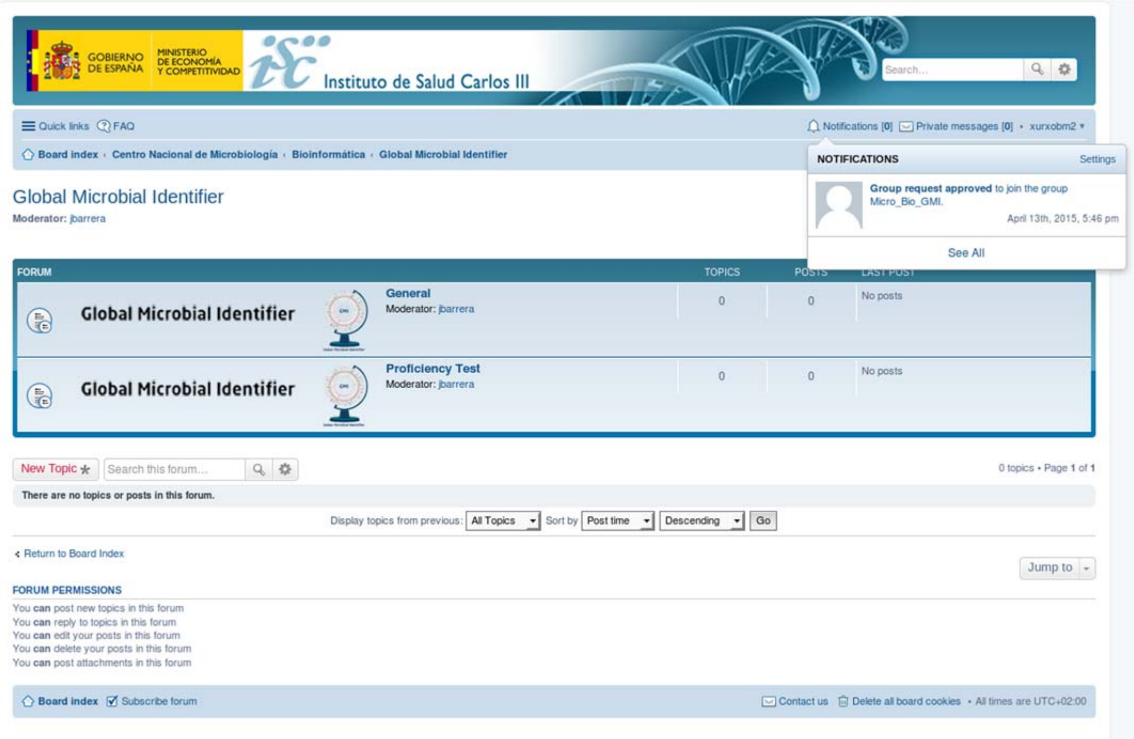
GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
Instituto de Salud Carlos III
Search…
Quick links
FAQ
Notifications [0]
Private messages [0] • xurxobm2
Board index

Request to join group

Are you sure you want to request joining the selected group?

Yes
No
Board index
Contact us
Delete all board cookies
• All times are UTC+02:00

Your request must be approved by GMI PT forum moderator. It should take not more than a few days. You will see a notification in the upper right corner of the screen when your request is approved.



GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
Instituto de Salud Carlos III
Search…
Quick links
FAQ
Notifications [0]
Private messages [0] • xurxobm2
Board index ‹ Centro Nacional de Microbiología ‹ Bioinformática ‹ Global Microbial Identifier

NOTIFICATIONS Settings

Group request approved to join the group Micro_Bio_GMI.
April 13th, 2015, 5:46 pm

See All

Global Microbial Identifier
Moderator: jbarrera

FORUM	TOPICS	POSTS	LAST POST
Global Microbial Identifier General Moderator: jbarrera	0	0	No posts
Global Microbial Identifier Proficiency Test Moderator: jbarrera	0	0	No posts

New Topic
Search this forum…

0 topics • Page 1 of 1

There are no topics or posts in this forum.

Display topics from previous: All Topics Sort by Post time Descending Go

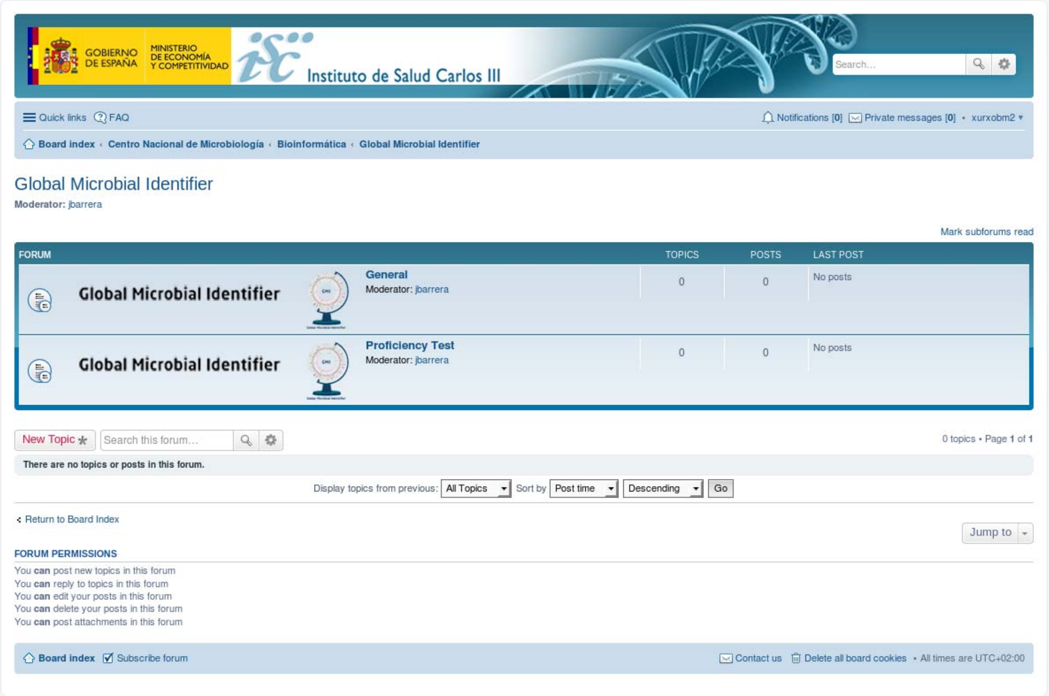
FORUM PERMISSIONS

You **can** post new topics in this forum
You **can** reply to topics in this forum
You **can** edit your posts in this forum
You **can** delete your posts in this forum
You **can** post attachments in this forum

Introduce yourself to other GMI PT participants

The first time you log into the forum after gaining admission to the *Micro_Bio_GMI* group were accepted you should introduce yourself. This is not mandatory but is recommended to facilitate communication among all GMI PT participants.

Click on ***General*** category link (if you do not see *Global Microbial Identifier* forum categories, see Troubleshooting, page 12).



Board index ‹ Centro Nacional de Microbiología ‹ Bioinformática ‹ Global Microbial Identifier

Global Microbial Identifier

Moderator: jbarrera

Mark subforums read

FORUM	TOPICS	POSTS	LAST POST
Global Microbial Identifier General Moderator: jbarrera	0	0	No posts
Global Microbial Identifier Proficiency Test Moderator: jbarrera	0	0	No posts

0 topics • Page 1 of 1

There are no topics or posts in this forum.

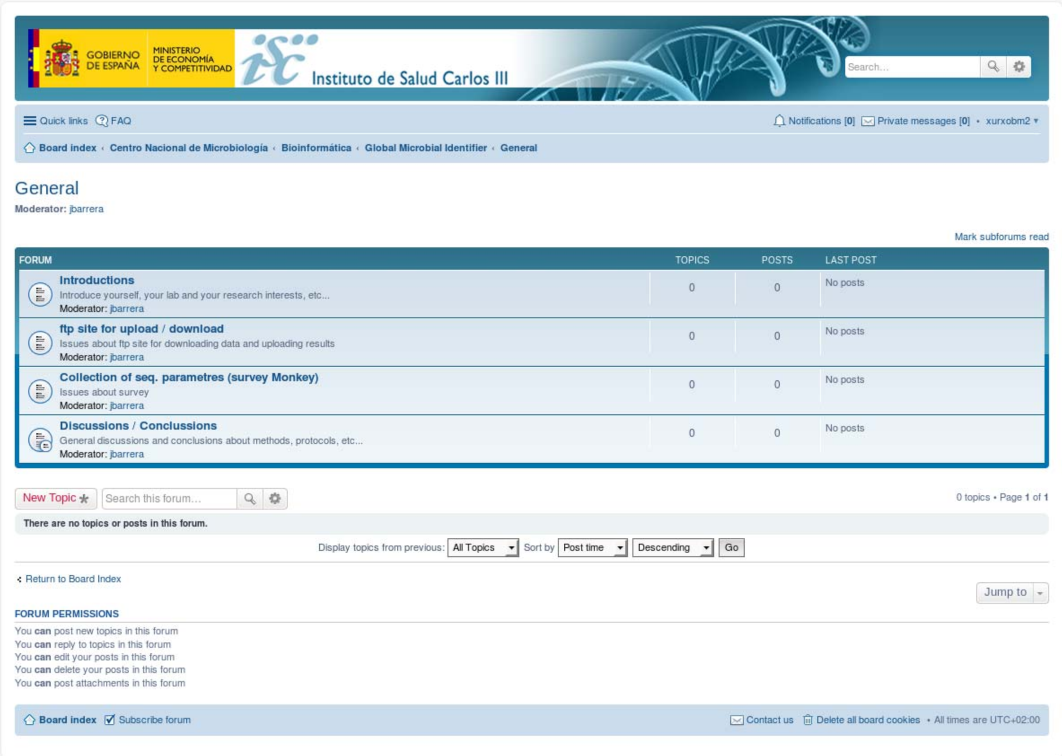
Display topics from previous: All Topics Sort by Post time Descending Go

FORUM PERMISSIONS

You **can** post new topics in this forum
You **can** reply to topics in this forum
You **can** edit your posts in this forum
You **can** delete your posts in this forum
You **can** post attachments in this forum

All times are UTC+02:00

Click on ***Introductions*** category link.



General

Moderator: jbarrera

FORUM	TOPICS	POSTS	LAST POST
Introductions Introduce yourself, your lab and your research interests, etc... Moderator: jbarrera	0	0	No posts
ftp site for upload / download Issues about ftp site for downloading data and uploading results Moderator: jbarrera	0	0	No posts
Collection of seq. parametres (survey Monkey) Issues about survey Moderator: jbarrera	0	0	No posts
Discussions / Conclussions General discussions and conclusions about methods, protocols, etc... Moderator: jbarrera	0	0	No posts

0 topics • Page 1 of 1

There are no topics or posts in this forum.

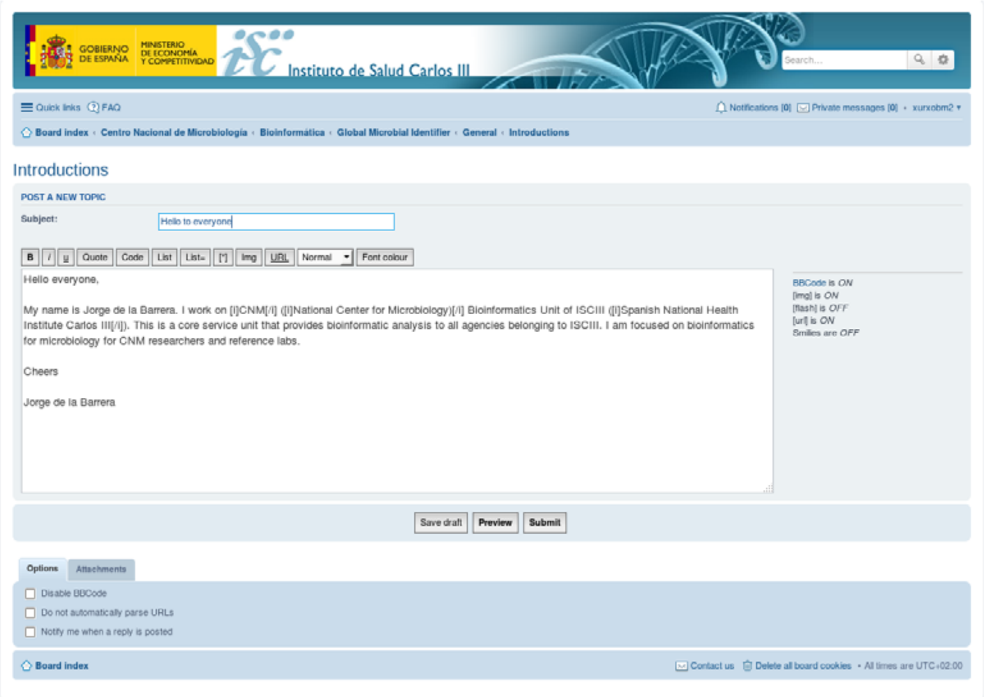
Display topics from previous: All Topics Sort by Post time Descending Go

‹ Return to Board Index

FORUM PERMISSIONS

You **can** post new topics in this forum
You **can** reply to topics in this forum
You **can** edit your posts in this forum
You **can** delete your posts in this forum
You **can** post attachments in this forum

Write a post introducing yourself and your lab and press ***Submit*** to publish it.



Introductions

POST A NEW TOPIC

Subject: Hello to everyone

Hello everyone,

My name is Jorge de la Barrera. I work on [i]CNM[/i] ([i]National Center for Microbiology)[/i] Bioinformatics Unit of ISCIII ([i]Spanish National Health Institute Carlos III[/i]). This is a core service unit that provides bioinformatic analysis to all agencies belonging to ISCIII. I am focused on bioinformatics for microbiology for CNM researchers and reference labs.

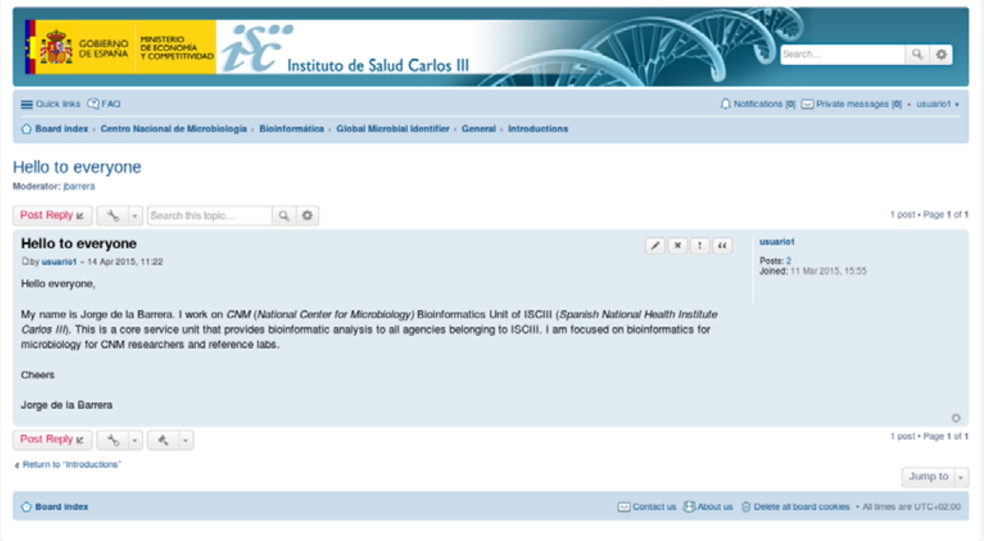
Cheers

Jorge de la Barrera

BBCode is *ON*
[img] is *ON*
[flash] is *OFF*
[url] is *ON*
Smilies are *OFF*

□ Disable BBCode
□ Do not automatically parse URLs
□ Notify me when a reply is posted

Once post is published it can be read by all GMI PT participants within ***Introductions*** forum category.



Hello to everyone

Moderator: jbarrera

Hello to everyone

by usuario1 » 14 Apr 2015, 11:22

Hello everyone,

My name is Jorge de la Barrera. I work on *CNM* (*National Center for Microbiology)* Bioinformatics Unit of ISCIII (*Spanish National Health Institute Carlos III*). This is a core service unit that provides bioinformatic analysis to all agencies belonging to ISCIII. I am focused on bioinformatics for microbiology for CNM researchers and reference labs.

Cheers

Jorge de la Barrera

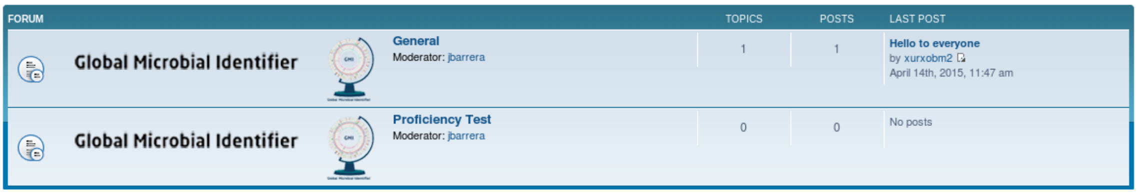
usuario1

Posts: 2
Joined: 11 Mar 2015, 15:55

Post a message on GMI PT forum

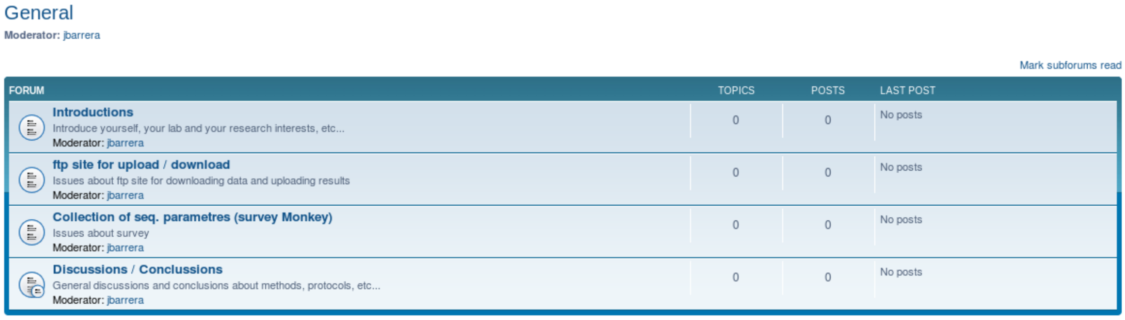
GMI PT forum is organized in categories to improve readability. When you want to write a message you should choose the right forum category to post on. This section details GMI PT forum categories.

Forum categories are organized in two branches: *General* and *Proficiency Test*. No messages can be posted at the main level.



FORUM		TOPICS	POSTS	LAST POST
Global Microbial Identifier	General Moderator: jbarrera	1	1	Hello to everyone by xurxobm2 April 14th, 2015, 11:47 am
Global Microbial Identifier	Proficiency Test Moderator: jbarrera	0	0	No posts

Introduction posts, issues about ftp site or survey and discussions/conclusions must be posted in the suitable category under *General* branch.



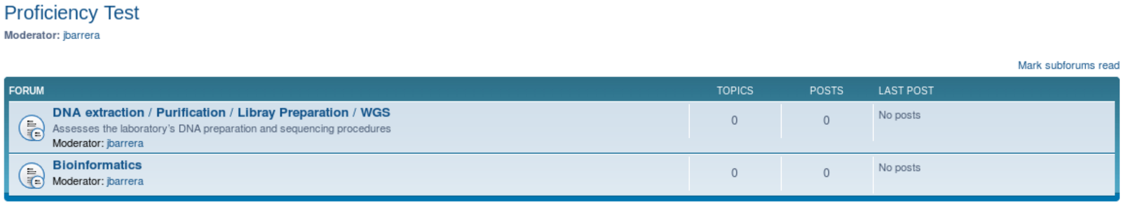
General

Moderator: jbarrera

Mark subforums read

FORUM	TOPICS	POSTS	LAST POST
Introductions Introduce yourself, your lab and your research interests, etc... Moderator: jbarrera	0	0	No posts
ftp site for upload / download Issues about ftp site for downloading data and uploading results Moderator: jbarrera	0	0	No posts
Collection of seq. parametres (survey Monkey) Issues about survey Moderator: jbarrera	0	0	No posts
Discussions / Conclussions General discussions and conclusions about methods, protocols, etc... Moderator: jbarrera	0	0	No posts

Profiency Test branch encompass categories related to the PT itself.



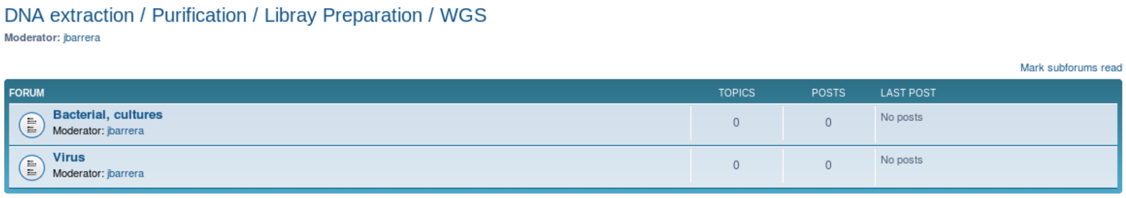
Proficiency Test

Moderator: jbarrera

Mark subforums read

FORUM	TOPICS	POSTS	LAST POST
DNA extraction / Purification / Libray Preparation / WGS Assesses the laboratory's DNA preparation and sequencing procedures Moderator: jbarrera	0	0	No posts
Bioinformatics Moderator: jbarrera	0	0	No posts

Wet lab and whole genome sequencing issues must be posted in the suitable category under *DNA extraction / Purification / Library Preparation / WGS* category.



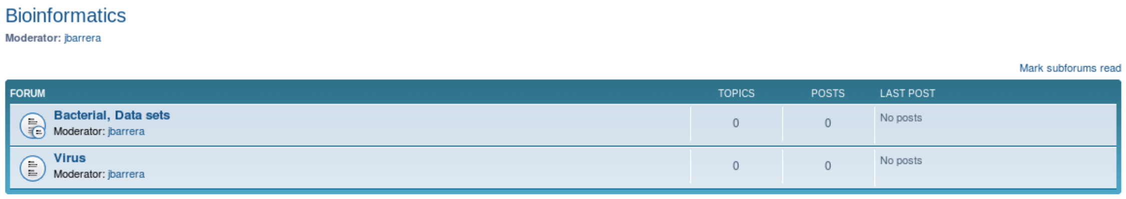
DNA extraction / Purification / Libray Preparation / WGS

Moderator: jbarrera

Mark subforums read

FORUM	TOPICS	POSTS	LAST POST
Bacterial, cultures Moderator: jbarrera	0	0	No posts
Virus Moderator: jbarrera	0	0	No posts

Dry lab issues must be posted in the suitable category under *Bioinformatics*.



Bioinformatics

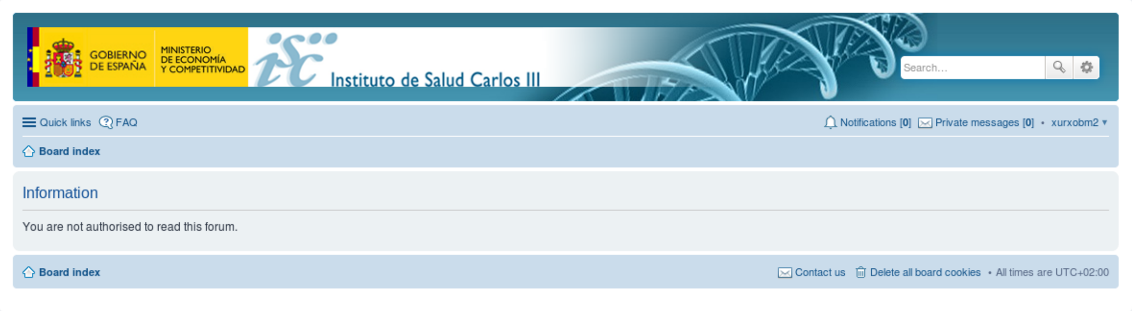
Moderator: jbarrera

Mark subforums read

FORUM	TOPICS	POSTS	LAST POST
Bacterial, Data sets Moderator: jbarrera	0	0	No posts
Virus Moderator: jbarrera	0	0	No posts

Troubleshooting

Problem: You are not authorized to read this forum



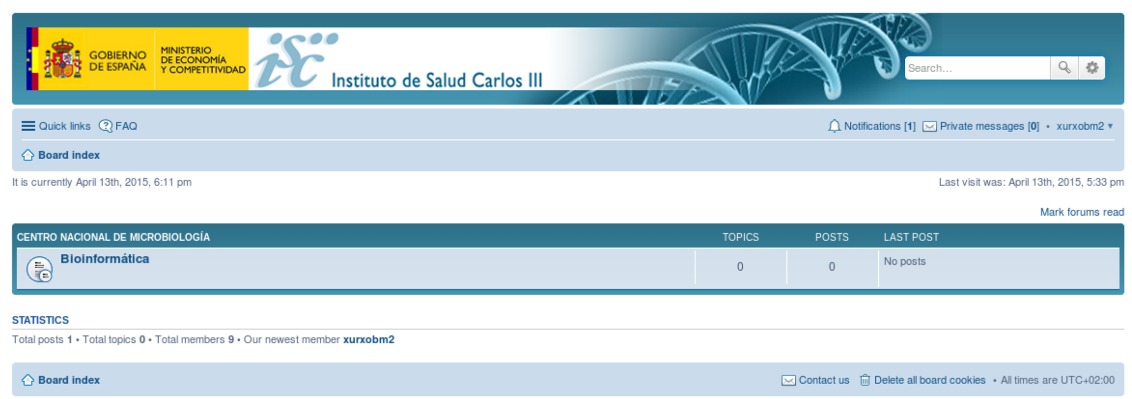
Information

You are not authorised to read this forum.

This message is shown if you have not registered.

Solution: Register as user following steps described in this guide.

Problem: Global Microbial Identifier forum categories are not shown



It is currently April 13th, 2015, 6:11 pm

Last visit was: April 13th, 2015, 5:33 pm

CENTRO NACIONAL DE MICROBIOLOGÍA	TOPICS	POSTS	LAST POST
Bioinformática	0	0	No posts

STATISTICS

Total posts **1** • Total topics **0** • Total members **9** • Our newest member **xurxobm2**

GMT PT forum is hosted by corporate forum site of *Instituto de Salud Carlos III*. Therefore, GMI PT forum is not the root forum. To access directly to GMI PT forum you should browse through *Bioinformática -> Global Microbial Identifier* or use URL *https://foros.isciii.es/viewforum.php?f=7*

GMI Proficiency Test, 2016

LabID:
Country:
Institute:
Main contact:
NGS contact (wet-lab):
Bioinformatics contact (dry lab):

Kgs. Lyngby, Denmark, October 2016

Dear [name],

Please find enclosed the bacterial cultures and DNA for the GMI Proficiency Test (PT), 2016.

Enclosed bacterial cultures and DNA

Depending on the level of your registration, the following live bacterial cultures are enclosed:

- codes GMI16-001-BACT and GMI16-002-BACT (2 *Campylobacter* strains)
- codes GMI16-003-BACT and GMI16-004-BACT (2 *Listeria monocytogenes* strains)
- codes GMI16-005-BACT and GMI16-006-BACT (2 *Klebsiella pneumoniae* strains)

The live bacterial cultures are lyophilized and sent as KWIK-STIKs™; each KWIK-STIK™ device features a single microorganism strain in a lyophilized pellet, a reservoir of hydrating fluid, and an inoculating swab (for more information, see the GMI PT protocol and http://microbiologics.com/s.nl/sc.7/category.98564/.f)

In addition, purified DNA corresponding to the bacterial cultures are enclosed:

- codes GMI16-001-DNA and GMI16-002-DNA (purified DNA from 2 *Campylobacter* strains)
- codes GMI16-003-DNA and GMI16-004-DNA (purified DNA from 2 *L. monocytogenes* strains)
- codes GMI16-005-DNA and GMI16-006-DNA (purified DNA from 2 *K. pneumoniae* strains)

The bacterial DNA is shipped as dried samples using a DNA stabilizing agent (DNAstable® *Plus*, Biomatrica).

Storage until handling

On arrival, The KWIK-STIKs™ must be refrigerated until handling in the laboratory.

On arrival, either rehydrate your DNA samples 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, or
(c) if sequencing the samples is planned within the first 10 days of arrival of the shipment, you may store the dried samples in the zip-lock bag in which they arrived along with the silica gel desiccant pack. Should moisture start to appear, the desiccant pack must be changed.

Usernames and passwords for download of sequences (for item 2; analysis of the provided datasets) and upload of results according to the description in the protocol:

For download:
Your username: [User_download]
Your password: [Password_download]

For upload:
Your username: [User_upload]
Your password: [User_upload]

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

Further information

On the GMI website, you find further information relevant for the GMI Proficiency Test (see http://www.globalmicrobialidentifier.org/Workgroups/About-the-GMI-Proficiency-Test-2016).

The protocol gives detailed descriptions for the testing the bacterial cultures (item 1a), the corresponding DNA (item 1b), and the analysis of the provided datasets (item 2). 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 for item 2, you need the username and password listed above.

Note the description in the protocol (in particular Appendix 4) of the web-based discussion forum available for participants in the GMI PT 2016.

Note that results must be submitted electronically no later than December 14th 2016.

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

Do not hesitate to contact us for further information,

Susanne Karlsmose Pedersen
GMI Proficiency Test Coordinator

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: 978-87-93565-14-2