SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Gen-FS coordinated proficiency test data for genomic foodborne pathogen surveillance, 2017 and 2018 exercises

Ruth E. Timme[1]✉, Patricia C. Lafon[4], Maria Balkey[1], Jennifer K. Adams[3], Darlene Wagner[4], Heather Carleton[2], Errol Strain[1], Maria Hoffmann[1], Ashley Sabol[2], Hugh Rand[1], Rebecca Lindsey[2], Deborah Sheehan[3], Joseph D. Baugher[1] & Eija Trees[3]

The US PulseNet and GenomeTrakr laboratory networks work together within the Genomics for Food Safety (Gen-FS) consortium to collect and analyze genomic data for foodborne pathogen surveillance (species include *Salmonella enterica*, *Listeria monocytogenes*, *Escherichia coli* (STECs), and *Campylobactor*). In 2017 these two laboratory networks started harmonizing their respective proficiency test exercises, agreeing on distributing a single strain-set and following the same standard operating procedure (SOP) for genomic data collection, running a jointly coordinated annual proficiency test exercise. In this data release we are publishing the reference genomes and raw data submissions for the 2017 and 2018 proficiency test exercises.

Background & Summary

Genomic surveillance of foodborne pathogens in the United States (US) is coordinated by two networks: US PulseNet[1,2] and GenomeTrakr[3]. PulseNet includes certified public health labs in all 50 US states; its mission is to collect and sequence isolates from human foodborne illnesses and identify outbreaks. The GenomeTrakr network consists of academic, agricultural, international, private, and public health labs; its mission is to sequence the isolates found in food and environmental sources (facility inspections, animals, water, etc), which provides leads for identifying sources of foodborne contamination events. Both networks submit their raw sequence data to the Sequence Read Archive (SRA) database at the National Center of Biotechnology Information (NCBI). These data[4,5] are also made available within NCBI's Pathogen Detection portal, which provides updated clustering of related isolates and information about antimicrobial resistance. The procedures for collecting and analyzing these whole genome sequencing (WGS) data are strictly governed by a multi-agency consortium called Genomics for Food Safety (Gen-FS)[6,7]. Members include the Centers of Disease Control and Prevention (CDC), the Food and Drug Administration (FDA), the NCBI, the United States Department of Agriculture (USDA) Food Safety Inspection Service (FSIS), Association of Public Health Laboratories (APHL), and state public health labs. Within Gen-FS, working groups harmonize lab protocols, standardize data quality requirements, verify new technologies, and coordinate an annual proficiency test (PT), among other activities. The PT exercises for PulseNet and GenomeTrakr were held separately until 2017, when the two networks agreed to hold a single annual PT exercise spanning both networks, harmonizing the strain-set and general SOP for data collection.

Each lab participating in PulseNet or GenomeTrakr networks must pass the annual PT to ensure proficiency in collecting WGS data for foodborne pathogens. In addition, these PT exercises are crucial for Gen-FS in accessing overall quality assurance (QA) and quality control (QC) for public health data collection. This exercise helps the Gen-FS coordinating team identify areas for improvement in sequence quality, data transfer, adherence to the SOP, QC thesholds, and communications. For PulseNet participating laboratories, passing the annual PT is a requirement for maintaining the laboratory's formal certification status which allows the laboratory to submit data to PulseNet's databases. If the initial PT submission fails, a laboratory is allowed to resubmit twice before

[1]US Food and Drug Administration, College Park, MD, USA. [2]Centers for Disease Control and Prevention, Atlanta, GA, USA. [3]Association of Public Health Laboratories, Silver Spring, MD, USA. [4]International Health Resource Co., Atlanta, GA, USA. ✉e-mail: ruth.timme@fda.hhs.gov

losing its certification. For GenomeTrakr participating laboratories, a successful PT submission qualifies as a "pass" (see Technical Verification for detail on QC thresholds used for inclusion) and resulting reports provide a rigerious statistical accessment of the submission. Previously published PT exercises[8,9] for WGS in foodborne pathogens were able to identify several interesting patterns including a low-rate of genetic variation among clonal isolates and stayed within expected error rates for several key areas, such as sequence quality, read mapping, assembly, insert sizes, and variant detection. Data also showed remarkable uniformity across dozens of laboratories.

Here we present the raw PT data collected from the Gen-FS PT exercises held in 2017 and 2018. For each year's PT exercise each participating laboratory received the same set of six isolates to sequence according to the Gen-FS harmonized SOP (the isolate sets were different each year). After following the data collection protocol, the labs submitted raw WGS data for each of the six isolates to their respective coordinating team for a full PT analysis (PulseNet, GenomeTrakr, or both). Coordinating teams then reported back to each participating laboratory with a "pass/fail" and an analysis report of the submitted data. The 12 laboratories that were members of both GenomeTrakr and PulseNet received two independently-generated reports back from each respective coordinating team.

For past technologies (e.g. PFGE), the general public or industry stakeholders would have had to issue a Freedom of Information Act (FOIA) request to access data collected from this type of exercise. However, in following our open data commitment across our foodborne pathogen surveillance effort, we are also releasing the data for our PT exercises. The raw sequence data collected, along with closed reference genomes for each of the isolates (6 for each year; total 12) were made public at NCBI.

As new chemistries are adopted within our established WGS workflow and new laboratories join the network, annual PTs play an important role for monitoring consistency while highlighting potential areas of improvement. Combing through these datasets each year enables us to understand laboratory-to-laboratory variation, to provide checks for our QA/QC thresholds, and to ensure proper verification for new chemistries, protocol changes, and new next generation sequencing (NGS) technologies as they come online. These quality assurance steps are important for public health and disease surveillance, and they also provide important transparency for the industries that are most effected by regulatory action (recalls, seizures, injunctions, etc.).

Methods

Reference genomes. A different set of strains was chosen for each PT exercise: four *Salmonella enterica* and two Shiga toxin producing *Escherichia coli* (STEC) strains were selected in 2017, and in 2018 four *S. enterica* and two *Listeria monocytogenes* strains were selected. Each of these 12 strains were closed on the Pacific Biosciences (PacBio) *RS II* sequencing platform to provide a baseline to which the results of the participating labs would be compared. The preparation, sequencing of the 20 kb libraries and the subsequent sequence analyses were carried out as described in Timme, R. E. *et al.*[8]. In summary, the libraries were prepared based on the 20 kb PacBio sample preparation protocol. Afterwards the libraries were sequenced using the P6/C4 chemistry on two to three single-molecule real-time (SMRT) cells (with size selection and without) with a 240-min collection time on the Pacific Biosciences *RS II* platform. Analysis of the continuous long reads was implemented using SMRT Analysis 2.3.0. and *de novo* assembly was performed using PacBio hierarchical genome assembly process HGAP 3.0[10] with default parameters. Resulting assemblies for both chromosomes and plasmids were checked manually for even sequencing coverage and were processed using Gepard[11] to identify overlapping regions at the ends. The improved consensus sequence was uploaded in SMRT Analysis 2.3.0 to determine the final consensus and accuracy scores using the Quiver consensus algorithm. Further, potential SNPs/indels were corrected with Pilon v1.18[12] using paired-end short-read data obtained from the Illumina MiSeq platform then mapped to the reference sequences via Bowtie2 v2.2.9[13].

Strain distribution. For each PT exercise, participating laboratories each recieved six lyophilized strains. In 2017, 64 laboratories participated; 12 of these labs were participants of both GenomeTrakr and PulseNet. In 2018, 78 laboratories participated, 13 of these labs were participants of both GenomeTrakr and PulseNet. (Online-only Table 1).

Strain revival for *S. enterica*, *E. coli* and *L. monocytogenes*. The lyophilized cells were resuspended with 1.0 mL sterile reagent grade water or trypticase soy broth (TSB), small amount was inoculated on a blood agar plate (BAP), and incubated overnight at 37 °C. A single, isolated colony was picked and streaked on fresh BAP and incubated at 37 °C overnight in aerobic conditions. The growth from this plate was used to make DNA templates. If no growth occurred on the initial plate, a second attempt was made by re-plating a larger volume of the resuspension.

DNA extraction, library prep and sequencing. Participating laboratories were instructed to use the Gen-FS harmonized SOP for DNA extraction, library preparation and DNA sequencing. DNA was extracted using Qiagen DNeasy (Qiagen, Hilden, Germany) kits. The libraries were prepared using the Nextera XT (Illumina, San Diego, CA) DNA library prep kit with one of two options: (a) the standard Illumina bead-based normalization or (b) manual normalization using library concentrations and estimated genome size. Sequencing was preformed using MiSeq Reagent Kit v2 (Illumina, San Diego, CA) chemistry for 2 × 250 cycles. Each PT run contained exactly 16 isolates: the 6 PT isolates along with 10 additional routine and/or historical isolates (replicates of the PT isolates were allowed). Participants populated the sequencing sample sheets with the following IDs: "Sample_ID" included the PulseNet proficiency identifiers and technician's initials; "Sample_name" included the sample ID, Lab ID, machine ID and run date (e.g. SAP18-8999jk-GA-M0947-180215.), and "Project", which was only used in the GenomeTrakr workflow to create unique folders within BaseSpace for data transfer.

Data transfer. While the strains and data collection were harmonized across this exercise, the data transfer, analysis and reporting were performed separately within PulseNet and GenomeTrakr for their own member laboratories. Depending on the type of laboratory and their access to data transfer services, there were several possible routes for transferring data. Laboratories with network access to BaseSpace Sequence Hub (Illumina) streamed their sequencing run(s) directly to BaseSpace, then shared their data with their respective network coordinating team(s). Non-federal laboratories without BaseSpace access transferred their raw data (as FASTQ files) through a secure file transfer protocol (SFTP) site. Laboratories within the federal network transferred their runs to an accessible shared drive. Along with the FASTQ files, each laboratory specified which variation in the library prep they followed: (a) the standard Illumina bead-based normalization or (b) manual normalization using library concentrations and estimated genome size.

Data Records

A single umbrella bioproject, PRJNA504454, and two data bioprojects for 2017[14] and 2018[15] were established at NCBI to hold all the data associated with this exercise. Each data BioProject contains six biosamples, describing the metadata for each of the six strains distributed during the respective PT exercises (Online-only Table 2). Complete reference genomes for each distributed strain (annotated, closed assemblies) were submitted to NCBI's Genbank (Online-only Table 2).The FASTQ files (raw sequence data) from each participating laboratory were submitted to NCBI's sequence read archive (SRA) database and linked to the appropriate biosample and bioproject (Supplemental File 1). In observing the norm of publishing proficiency test results, we have de-identifed the laboratories from their respective data submissions[16–21]. The individual PT evaluation results for each laboratory are confidential. However, we are extending beyond the norm by releasing all the names of participating laboratories (Online Table 1) *and* the raw data collected across the exercise. PT exercise results are a snapshot in time that may or may not reflect deeper quality control issues in a laboratory. Coordinating bodies worked directly with laboratories that underperformed, identifying and solving any QC issues that appeared systemic. Our goal with this data release is to communicate the value of the entire dataset without fears of public or legal retribution for the participants.

Technical Verification

Internal QC to determine validity of data for both 2017 and 2018. In order to ensure maximal utility, the dataset was restricted to samples that passed a series of QC thresholds set by each network. Although the isolates and sequencing protocols were harmonized across the PT exercises, each coordinating body ran their own analyses and distributed their own reports, reflecting each bodies different goals for the exercise (e.g. PulseNet used graded reports because many PN labs need them for accreditation purposes (e.g. CLIA) and GenomeTrakr used a statistical assessment-style report with the goal of using the assessments to identify problem areas and improve overall quality).

PulseNet utilized standardized organism-specific evaluation forms and a grading system to evaluate the PT submissions for critical and non-critical quality metrics (Supplemetal Files 2 and 4). Failing to meet the minimum threshold/acceptable range for any of the critical quality metrics resulted in an automatic failure, while failing to meet the minimum threshold for the non-critical metric (insert size) resulted in points deduction. A maximum of 100 points could be accumulated and a minimum of 85 points was required for passing. The following rejection criteria were used for the critical quality metrics:

1. Average coverage $<$ 20x for *Listeria*, $<$ 30x for *Salmonella*, and $<$ 40x for *Escherichia*
2. Read 1 and Read 2 average Q score $<$ 28.00. Sequences with quality scores 28.00–29.99 were accepted with 10–20x additional coverage but resulted in points deduction
3. Assembled genome size $>$ 5% outside the expected size
4. Percentage of core genome genes detected $<$ 95% (*Listeria* only)
5. Number of hqSNP differences $>$ 1 per megabase compared to the reference sequence
6. Number of cgMLST allele differences $>$3 compared to the reference sequence

Analysis summary for PulseNet. In 2017, of the 31 participating laboratories, 30 passed the PT for *Salmonella* and STEC with an average passing score of 99 for each organism (Supplemental File 2). In 2018, of the 46 participating laboratories, 42 passed for *Listeria* with an average passing score of 96 and 40 passed for *Salmonella* with an average score of 97 (Supplemental File 4). The failed submissions were caused by low average coverage or quality score, incorrect genome size, low percentage of core genome detected and apparent isolate mix-ups as evidenced by high allele and hqSNP differences compared to the reference sequence. Many labs also appeared to struggle in meeting the minimum insert size of 300 bp. No laboratories lost their certification status as resubmissions were successful. As a follow-up to the detected insert size problem, PulseNet developed focused troubleshooting and training materials materials on improving the insert length and recommended that the laboratories switch from the Nextera XT library preparation kit to the newer Nextera DNAFlex kit.

GenomeTrakr excluded any individual samples failing to meet the minimal expected thresholds for average coverage ($<$20X) and average read quality (Q score $<$ 28.00). Entire sequencing runs were excluded (equivilant to a PulseNet failure) for the following reasons:

1. Any evidence of sample misannotation
2. Any evidence of noncompliance with the SOP (including read length, library prep kit, sequencing chemistry, and number of samples per run)
3. Runs which included too few acceptable PT samples ($<$4).

Submissions that passed this initial screening were included in the exercise and a statistical report was generated placing each PT submission in context with data from the entire exercise (example reports included in Supplemental Files 3 and 5). Labs were given the option to re-sequence the panel if their submission included QC thresholds far outside the normal range (lower and upper quartiles) in an effort to encourage the labs to make improvements within their laboratory workflows based on feedback from their PT assessment.

Usage Notes

We see many possible uses for this dataset, especially to demonstrate the reliability of sequencing data produced by large networks of diverse laboratories. We encourage the use of this dataset for competency assessments, verifying or validating new chemistries and platforms, and PT:

1. Obtain and sequence, while following the Gen-FS SOP, a subset of the PT isolates described in this manuscript (for access to the strains, please contact PulsenetNGSlab@cdc.gov - the ATCC catalogue numbers for the strains were pending at the time of the publication),
2. Using a bioinformatics analysis pipeline of choice, analyze the new sequencing data along with the relevant subset of data described in this manuscript (SRA run accessions for downloading data listed in Supplemental File 1),
3. For each desired analytical metric, assess the new data as part of a distribution representing all of the available runs of the same isolate.

Importantly, in contrast to single threshold-based assessments, these data enable individual assessments using standard statistical methods such as interquartile range and outlier detection (see example statistical analyses in Supplemental Files 3 and 5). This allows laboratories to gain the feedback of participating in large-scale PT exercise without having to participate directly.

Received: 28 February 2020; Accepted: 20 October 2020;
Published online: 19 November 2020

References

1. Nadon, C. *et al*. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. *Euro Surveill.* **22**, 30544 (2017).
2. Swaminathan, B., Barrett, T. J., Hunter, S. B., Tauxe, R. V. & PulseNet, C. D. C. Task Force. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerg. Infect. Dis.* **7**, 382–389 (2001).
3. Allard, M. W. *et al*. Practical Value of Food Pathogen Traceability through Building a Whole-Genome Sequencing Network and Database. *J. Clin. Microbiol.* **54**, 1975–1983 (2016).
4. NCBI Resource Coordinators. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **45**, D12–D17 (2017).
5. NCBI. NCBI Pathogen Detection. *NCBI Pathogen Detection* https://www.ncbi.nlm.nih.gov/pathogens/.
6. CDC. *Annual Report to the Secretary, Department of Health and Human Services*. 22–23 (2015).
7. Brown, E., Dessai, U., McGarry, S. & Gerner-Smidt, P. Use of Whole-Genome Sequencing for Food Safety and Public Health in the United States. *Foodborne. Pathog. Dis.* **16**, 441–450 (2019).
8. Timme, R. E. *et al*. GenomeTrakr proficiency testing for foodborne pathogen surveillance: an exercise from 2015. *Microb. Genom.* **57**, 289 (2018).
9. Moran-Gilad, J. *et al*. Proficiency testing for bacterial whole genome sequencing: an end-user survey of current capabilities, requirements and priorities. *BMC Infect. Dis.* **15**, 174 (2015).
10. Chin, C.-S. *et al*. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
11. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).
12. Walker, B. J. *et al*. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One* **9**, e112963 (2014).
13. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**(4), 357–359 (2012).
14. 2017 GenomeTrakr/PulseNet Proficiency Test Exercise. *NCBI BioProject* https://identifiers.org/ncbi/bioproject:PRJNA507262 (2018).
15. 2018 GenomeTrakr/PulseNet Proficiency Test Exercise. *NCBI BioProject* https://identifiers.org/ncbi/bioproject:PRJNA507264 (2018).
16. Davies, K. D. *et al*. Multi-Institutional FASTQ File Exchange as a Means of Proficiency Testing for Next-Generation Sequencing Bioinformatics and Variant Interpretation. *The Journal of Molecular Diagnostics* **18**, 572–579 (2016).
17. Zhong, Q. *et al*. Multi-laboratory proficiency testing of clinical cancer genomic profiling by next-generation sequencing. *Pathology - Research and Practice* **214**, 957–963 (2018).
18. Hegde, M. *et al*. Development and Validation of Clinical Whole-Exome and Whole-Genome Sequencing for Detection of Germline Variants in Inherited Disease. *Arch. Pathol. Lab. Med*. arpa.2016-0622-RA, https://doi.org/10.5858/arpa.2016-0622-RA (2017).
19. Feldman, G. L., Schrijver, I., Lyon, E., Palomaki, G. E. & CAP/ACMG Biochemical and Molecular Genetics Resource Committee. Results of the College of American Pathology/American College of Medical Genetics and Genomics external proficiency testing from 2006 to 2013 for three conditions prevalent in the Ashkenazi Jewish population. *Genet. Med*. **16**, 695–702 (2014).
20. Osoegawa, K. *et al*. Quality control project of NGS HLA genotyping for the 17th International HLA and Immunogenetics Workshop. *Human Immunology* **80**, 228–236 (2019).
21. Raggi, C. C., Pinzani, P., Paradiso, A., Pazzagli, M. & Orlando, C. External Quality Assurance Program for PCR Amplification of Genomic DNA: An Italian Experience. *Clin Chem* **49**, 782–791 (2003).

Acknowledgements

We'd like to thank the participating laboratories listed in Online-only Table 2. We'd also thank Lili Fox Vélez, Office of Regulatory Science, for scientific writing support.

Author contributions

Ruth E. Timme: responsible for the concept and design of the study, wrote and edited the paper. Patricia C. Lafon: responsible for implementation and distribution of PT, tested the candidate PT isolates prior to selection, performed data processing, sequencing analysis, data submission, wrote SOP. Maria Balkey: responsible for implementation and distribution of PT, performed data processing, data submission to NCBI, wrote SOP, and helped edit and revise the manuscript. Jennifer K. Adams: responsible for implementation and distribution of PT, performed data processing, sequencing analysis, data submission, distribution of results, wrote SOP and helped edit manuscript. Darlene Wagner: performed data processing, and sequencing analysis. Heather Carleton: helped edit and revise the manuscript. Errol Strain: helped edit and revise the manuscript. Maria Hoffmann: Performed data processing, and sequencing analysis Ashley Sabol: tested the candidate PT isolates prior to selection Hugh Rand: helped edit and revise the manuscript. Rebecca Lindsey: made the PT strains available at ATCC, helped edit and revise the manuscript. Deborah Sheehan: performed data processing. Joseph Baugher: performed data processing, sequencing analysis, and helped edit and revise the manuscript. Eija Trees: responsible for the concept and design of the study, wrote SOP, helped edit and revise the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41597-020-00740-7.

Correspondence and requests for materials should be addressed to R.E.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.