A Dynamic Whole-genome Database for Comparative Analyses, Molecular Epidemiology and Phenotypic Summary of Bacterial Pathogens

Chad R. Laing[1,2], Eduardo Taboada[1], Peter Kruczkiewicz[1,2], James E. Thomas[2] and Victor P. J. Gannon[1]

[1]*Laboratory for Foodborne Zoonoses, Public Health Agency of Canada, Lethbridge, Alberta, Canada*

[2]*Department of Biological Sciences, University of Lethbridge, Lethbridge, Alberta, Canada*

Keywords: Genomics, Database, Molecular Epidemiology, Phenotype, Comparative Analyses.

Abstract: Background. Recent outbreaks caused by bacterial contaminants in food, including sprouts by *E. coli* O104:H4 in Germany and processed meats by *Listeria* in Canada highlight the need for rapid and accurate characterization of bacterial pathogens. Current sequencing platforms have revolutionized the amount and quality of data available to epidemiologists, public health officials and microbiologists, who now require powerful yet intuitive tools to make sense of the underlying biology in these large datasets. In this study, we developed bioinformatics tools to: automate whole-genome analyses, make the data broadly accessible via novel reporting functions, and provide a dynamic computational platform for genomic analyses online at http://76.70.11.198/bacpath.

Methods. A PHP-based web front end and PostgreSQL database display the pre-computed data. Genomic comparisons are performed using updates to our previously created pan-genomic software suite, Panseq (http:://lfz.corefacility.ca/panseq/). New genomic sequences are analyzed and added to the database without the need for recomputing previous analyses. Phylogenetic trees are created with MrBayes. Statistical calculations are performed using R.

Results. A pathogen-specific genomic database encompassing all publicly available *E. coli* strains was created as a proof of concept. Pre-computed comparisons for the hundreds of bacterial genomes including phylogeny, presence/absence of virulence markers, group-specific biomarkers and geospatial information were generated. Data reporting tools were created to summarize the complexity of the data and to provide biologically pertinent results including genotype, phenotype (eg. anti-microbial resistance), and geospatial information.

Discussion. The database provides rapid and accurate identification and characterization of *E. coli*. Output is formatted specifically for end users describing virulence, phylogeny and group-specific markers. Uptake of a global surveillance system with near real time analysis will provide an effective early warning system and allow for a faster response to pathogen-related outbreaks.

1 INTRODUCTION

Recent outbreaks caused by bacterial contaminants in food, including sprouts by *E. coli* O104:H4 (Grad et al., 2012) in Germany and processed meats by *Listeria monocytogenes* in Canada (Gilmour et al., 2010) highlight the need for rapid and accurate identification and characterization of bacterial pathogens. Next generation nucleic acid sequencing platforms have revolutionized all areas of microbiology relevant to food safety and public health. There has been a relentless increase in the capacity, speed and portability of whole-genome sequencing (WGS) systems coupled with a steady decline in the cost of analysis, which has caused the amount and quality of data available to epidemiologists, public health officials and microbiologists to increase to discipline transforming levels.

We have already seen examples in epidemiology where the sources of food and water borne disease outbreaks such as the *Vibrio cholerae* outbreak in Haiti and the aforementioned *E. coli* O104:H4 (Grad et al., 2012) and *Listeria* (Gilmour et al., 2010) were determined using whole-genome sequencing. Within public health, identifying transmission routes within a

R. Laing C., Taboada E., Kruczkiewicz P., E. Thomas J. and P. J. Gannon V..
A Dynamic Whole-genome Database for Comparative Analyses, Molecular Epidemiology and Phenotypic Summary of Bacterial Pathogens.
DOI: 10.5220/0004239803040307
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2013), pages 304-307
ISBN: 978-989-8565-35-8

hospital was previously impossible for highly related bacteria such as methicillin-resistant *Staphylococcus aureus* (MRSA), due to the inability of standard clinical typing methods to discriminate among isolates. Recently, WGS was used to prevent continued transmission of MRSA within a neonatal intensive care unit by identifying transmission events in a clinically relevant time-frame (Kser et al., 2012). Many areas of microbiology have moved to a comparative genomics paradigm, including creating molecular typing schemes for pathogens such as *E. coli* O157 and *Campylobacter jejuni* (Laing et al., 2008; Taboada, 2012).

With the data available and studies showing how powerful it can be if properly utilized, the only remaining hurdle is an easy-to-use system that allows epidemiologists, public health officials and microbiologists to access the data, tools for manipulating the data, and reporting functions, which provide the biological meaning to end users.

We have previously created a suite of pan-genomic analysis tools, Panseq (http://lfz.corefacility.ca/panseq/). In this study we designed and implemented an online computational platform that pre-computes analyses among thousands of bacterial genomes and efficiently analyzes new sequences by computing relationships within small strain clusters, allowing near real time comparisons among genomic sequences. Smart heuristic algorithms avoid the unnecessary recomputation of previously run analyses. As a proof of concept, a central genomics data repository with all publicly available genomic sequences for *E. coli* species was created. Additionally we implemented pre-computed analyses; automated calculation and real- time manipulation of phylogenies; the identification of genomic markers (presence/absence of specific genomic regions, and single-nucleotide polymorphisms); bio-statistical tools to find associations between group-specific genomic markers and phenotypic metadata (e.g., geospatial distribution, host, source); automated *in silico* genotyping (e.g., multi-locus sequence typing, antimicrobial resistance and virulence markers via the MIST platform); and reporting of interpreted data for use by microbial ecologists, physiologists, clinical researchers and epidemiologists.

2 METHODS

2.1 Database Setup

A webserver programmed in PHP using the Smarty template engine (http://smarty.incutio.com) allows access to and displays the pre-computed analyses stored in a PostgreSQL database v9.2.1 (http://www.postgresql.org/), which is highly scalable and fast (up to 350000 read queries per second, supporting up to 64 computing cores). The software is hosted on a linux CentOS server with 16 CPUs and 64 GB of RAM. The geographical mapping uses the Google Maps API (https://developers.google.com/maps/). All users may perform analyses and download data; however, only registered users may submit data for inclusion in the database. An overview of the analysis platform is shown in Figure 1.

2.2 Comparative Genomics

Bacterial comparisons including the presence / absence of virulence factors and antimicrobial resistance genes, segmentation and alignment of whole-genomes, and the determination of group-dominant loci are performed using updates to our previously created pan-genomic software suite, Panseq (Laing et al., 2010; Laing et al., 2011). New genomic sequences are analyzed and added to the database without the need for recomputing previous analyses.

2.3 Phylogeney

The fragmented and aligned genomes are generated using the Panseq core / accessory module with the following settings: minimumNovelRegionSize = 500, fragmentationSize = 500, nucB = 200, nucC = 50, nucD = 0.12, nucG = 100, nucL = 20, percentIdentityCutoff = 85, coreGenomeThreshold = 3. Phylogenetic trees are created using MrBayes (Ronquist and Huelsenbeck, 2003). Phylogenetic tree manipulations are performed in real time using The Newick Utilities (Junier and Zdobnov, 2010).

2.4 *In-silico* Typing

In-silico genotyping will use the molecular *in-silico* typing (MIST) package developed using C# and the .NET framework v4, which uses Blast for sequence comparisons and generates multi-locus sequence typing profiles, multi-locus variable number tandem repeat analysis profiles and molecular serotype designa-

tions. These typing results will automatically be provided when a sequence is uploaded to the database.

3 DATABASE DESCRIPTION

3.1 Strain Information

The computational genomics platform (http://76.70.11.198/bacpath) provides access to a listing of all indexed strains, currently all *E. coli*, with the capability of expanding to include all bacterial pathogens of interest. Geographical information for each strain is provided on a worldwide zoomable map. Each strain also contains metadata for host, source, date of isolation, geographical location, serotype and genomic sequence data; missing information is denoted as 'unknown'. Additional metadata tags can be added as the system grows.

3.2 Virulence and AMR

In addition to strain information, virulence and AMR factors are also provided, with a description of the function of the factor, its sequence and sequence statistics, genome / plasmid location and a linked list of all strains the factor is present in. To investigate the presence of factors among database strains, a factor or group of factors is selected among a user-defined group of genomes. This can be achieved by using the list interface or by highlighting groups of strains via the map of the world interface. The resulting output is a table of selected stains and the presence / absence of selected factors, which can be downloaded in tab-delimited, HTML, or XML format. The selected strains are also displayed geographically on the map of the world with the option of highlighting via color change, strains that contain factors of interest.

3.3 Phylogeny

The phylogeny of all strains is pre-computed and can be manipulated in real-time by the user (Figure 4). By selecting strains of interest, only the phylogenetic relationship among the selected strains will be displayed. Selected phylogenies can be displayed in circular or orthogonal format and downloaded as either Newick or Nexus files.

3.4 Group Comparison

Identification of virulence and / or AMR factors, as well as the presence / absence of genomic regions and single-nucleotide polymorphisms that are statistically dominant for a group can also be performed. The user selects two groups via any of three interfaces: the worldwide map interface; the list of strains interface; or the phylogeny interface where nodes of the tree can be selected. The identification of loci that differ statistically between the two groups is performed using Fisher's Exact test and displayed in tabular format ranked by p-value. Results can be downloaded as a tab-delimited text file. When a comparison is run, the results are saved in the database for future instant recall, and computed by the server after the user makes her selection.

3.5 Updating

Users who register an account may upload their own sequence, which is processed and added to the database under three data release modes: private, which allows only the registered user to see the strain in the database; private until a specified date, after which point it becomes public and accessible by anyone using the database; and public, which allows the strain to be immediately accessible to all database users.

4 DISCUSSION

4.1 Benefit

The computational platform will provide geographical and temporal mapping capabilities for occurrence of specific pathogen genotypes at community, national and international levels. It will also provide genotype-phenotype linkages with clinical and epidemiological data such as outbreak associations, likely sources of food and water contamination, and disease progression and outcomes. An international bacterial pathogen database will also lead to more rapid and accurate identification and characterization of food and waterborne bacterial pathogens, with an emphasis on the most effective potential interventions, treatments and procedures.

It will enable rapid recognition of pathogen-related infections and outbreaks resulting in a faster response, identification of risk groups and possible pathogen sources, a decrease in the number of infections within the community and more effective food recalls. An international pathogen database will allow a more effective early warning system for remote jurisdictions and possibly lead to decreased infrastructure costs and reduced concern about food safety.

The bioinformatic tools that we have developed make use of pre-computed genomic analyses that will be able to accommodate continued influx of genomic sequence data, requiring only new genomic data to be analysed. The results of analyses using the database allow end users to easily identify whether an isolate is exceptionally virulent, or not usually associated with human infection based on the presence / absence of known virulence attributes and AMR genes, and genomic similarity to other known human pathogens.

4.2 Collaboration

Two similar efforts to construct genomic databases for molecular epidemiology have recently been proposed (Kupferschmidt, 2011; FDA, 2012). While strain transport between countries can be difficult or impossible, genomic sequence information can be transmitted instantly, allowing rapid analyses and potentially life-saving interventions. International agencies need to be willing to share information between databases or to collaborate in building a single, multi-national database to fully realize the potential of comparative genomics, and individual strains need to be analyzed in the context of as many similar strains as possible to put the data in the proper context.

ACKNOWLEDGEMENTS

We would like to thank the Canadian Food Inspection Agency for allowing this research to be conducted at the Animal Diseases Research Institute. This work was supported by the Public Health Agency of Canada and grants from the Natural Sciences and Engineering Research Council of Canada (www.nserc-crsng.gc.ca) and Alberta Innovates Technology Futures (www.albertatechfutures.ca).

REFERENCES

FDA (2012). Press announcements - FDA, UC davis, agilent technologies and CDC to create publicly available food pathogen genome database. http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm311661.htm. The U.S. Food and Drug Administration (FDA), the University of California, Davis, Agilent Technologies Inc., and the Centers for Disease Control and Prevention (CDC) announced today a collaboration to create a public database of 100,000 foodborne pathogen genomes to help speed identification of bacteria responsible for foodborne outbreaks.

Gilmour, M., Graham, M., Van Domselaar, G., Tyler, S., Kent, H., Trout-Yakel, K., Larios, O., Allen, V., Lee, B., and Nadon, C. (2010). High-throughput genome sequencing of two listeria monocytogenes clinical isolates during a large foodborne outbreak. *BMC Genomics*, 11(1):120.

Grad, Y. H., Lipsitch, M., Feldgarden, M., Arachchi, H. M., Cerqueira, G. C., Fitzgerald, M., Godfrey, P., Haas, B. J., Murphy, C. I., Russ, C., Sykes, S., Walker, B. J., Wortman, J. R., Young, S., Zeng, Q., Abouelleil, A., Bochicchio, J., Chauvin, S., Desmet, T., Gujja, S., McCowan, C., Montmayeur, A., Steelman, S., Frimodt-Mller, J., Petersen, A. M., Struve, C., Krogfelt, K. A., Bingen, E., Weill, F.-X., Lander, E. S., Nusbaum, C., Birren, B. W., Hung, D. T., and Hanage, W. P. (2012). Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in europe, 2011. *Proceedings of the National Academy of Sciences of the United States of America*, 109(8):3065–3070. PMID: 22315421.

Junier, T. and Zdobnov, E. M. (2010). The newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics (Oxford, England)*, 26(13):1669–1670. PMID: 20472542.

Kser, C. U., Holden, M. T. G., Ellington, M. J., Cartwright, E. J. P., Brown, N. M., Ogilvy-Stuart, A. L., Hsu, L. Y., Chewapreecha, C., Croucher, N. J., Harris, S. R., Sanders, M., Enright, M. C., Dougan, G., Bentley, S. D., Parkhill, J., Fraser, L. J., Betley, J. R., Schulz-Trieglaff, O. B., Smith, G. P., and Peacock, S. J. (2012). Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *The New England journal of medicine*, 366(24):2267–2275. PMID: 22693998.

Kupferschmidt, K. (2011). Outbreak detectives embrace the genome era. *Science*, 333(6051):1818–1819.

Laing, C., Buchanan, C., Taboada, E. N., Zhang, Y., Kropinski, A., Villegas, A., Thomas, J. E., and Gannon, V. P. J. (2010). Pan-genome sequence analysis using panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics*, 11:461. PMID: 20843356.

Laing, C., Pegg, C., Yawney, D., Ziebell, K., Steele, M., Johnson, R., Thomas, J. E., Taboada, E. N., Zhang, Y., and Gannon, V. P. J. (2008). Rapid determination of escherichia coli O157:H7 lineage types and molecular subtypes by using comparative genomic fingerprinting. *Applied and Environmental Microbiology*, 74(21):6606–15. PMID: 18791027.

Laing, C., Villegas, A., Taboada, E. N., Kropinski, A., Thomas, J. E., and Gannon, V. P. J. (2011). Identification of salmonella enterica species- and subgroup-specific genomic regions using panseq 2.0. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*. PMID: 22001825.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Taboada, E. N. e. a. (2012). Development and validation of a comparative genomic fingerprinting method for high-resolution genotyping of *Campylobacter jejuni*. *Journal of clinical microbiology*, 50(3):788–797. PMID: 22170908.