BioData Mining

SHORT REPORT Open Access

Updating microbial genomic sequences: improving accuracy & innovation

Hongseok Tae, Enusha Karunasena, Jasmin H Bavarva and Harold R Garner*

* Correspondence:
garner@vbi.vt.edu
Virginia Bioinformatics Institute at Virginia Polytechnic Institute and State University, 1015 Life Sciences Circle, Blacksburg, VA 24061, USA

Abstract

Background: Many bacterial genome sequences completed using the Sanger method may contain assembly errors due in-part to low sequence coverage driven by cost.

Findings: To illustrate the need for re-sequencing of pre-nextgen genomes and to validate sequenced genomes, we conducted a series of experiments, using high coverage sequencing data generated by a Illumina Miseq sequencer to sequence genomic DNAs of *Bacteroides fragilis* NCTC 9343, *Salmonella enterica* subsp. enterica serovar Paratyphi A str. ATCC 9150, *Vibrio cholerae* O1 biovar El Tor str. N16961, *Bacillus halodurans* C-125 and *Caulobacter crescentus* CB15, which had previously been sequenced by the Sanger method during the early 2000's.

Conclusions: This study revealed a number of discrepancies between the published assemblies and sequence read alignments for all five bacterial species, suggesting that the continued use of these error-containing genomes and their genetic information may contribute to false conclusions and/or incorrect future discoveries when they are used.

Keywords: Microbiota, Genomics, Genomes, Bacteria, Sequences, Salmonella, Mycobacterium, Brucella, Vibrio

Findings

The completed genome sequences of over 2,000 bacterial species have been published during the last decade and many of them (we estimate at least 500) were sequenced exclusively by the Sanger method; however this method was frequently deployed at low sequence coverage due to cost constraints. Even though the Sanger method assemblies targeted high accuracy (99.5%), low coverage might leave assembly errors in the completed genome sequences, which have been frequently used as references for re-sequencing projects. At the start of a re-sequencing analysis, it is important to choose a suitable reference genome sequence to compare against, to better identify high probability variants. These "variations" are then a foundation for many downstream correlative and functional analyses. Significantly, in the analysis of pathogens such as *Brucella*, *Salmonella* and *Vibrio* species, the results of variation detection are the basis for developing assays that are critical to the detection and validation of these pathogens.

In our previous work [1] with *Brucella suis* 1330, which was sequenced with the Sanger method in 2002 [2] and re-sequenced in 2011 using the Illumina GAIIx platform, we identified a number of discrepancies between the published and the new

BioMed Central

assembly. We used a hybrid approach of mapping and assembly with the Illumina sequencing data, and identified a total of twelve very high confidence sequence differences including ten INDELs (insertions or deletions) and two substitutions between the assemblies. Among them, six INDELs caused frameshifts within protein-coding loci. The differences were significant enough that the published sequence could lead downstream studies into inaccurate reporting and understanding of genomic mutations. Another re-sequencing study by Wynne [3] for the genome of *Mycobacterium avium* subsp. *paratuberculosis* K10 also showed differences between its original assembly and revised assembly which was originally sequenced in 2005 (Sanger method) and later with the Illumina GAIIx platform in 2010. Importantly, these studies implicate that other completed bacterial genome assemblies sequenced with the Sanger method may contain assembly errors resulting in inaccurate variation analyses. It also highlights the need for re-sequencing efforts using high coverage sequencing data generated by efficient and cost effective next-generation sequencing (NGS) technologies to validate these genome sequences. Especially for pathogen genomes, accurate references are essential for studying, detecting, and preventing public safety threats. Additionally, billions of dollars are invested by multiple federal agencies (i.e. CDC, FDA, USDA, and NIH) and private institutions (i.e. food production facilities, pharmaceutical companies, diagnostics labs etc.…), annually, to maintain safety from these biological agents; consequently, these efforts are now more frequently reliant upon standardized genomic information for genetic testing that utilize established markers for pathogen identification. Inaccurate or incomplete genomic information could contribute to misinformation to these agencies, impacting human health in addition to their effect on basic research.

Methods

To provide reliable supporting data for our observations, we sequenced five bacterial genomes of which sequences had been completely assembled and published in the early 2000's using the Sanger method. The five bacteria include *Bacteroides fragilis* NCTC 9343 [4], *Salmonella enterica* subsp. enterica serovar Paratyphi A str. ATCC 9150 [5], *Vibrio cholerae* O1 biovar El Tor str. N16961 [6], *Bacillus halodurans* C-125 [7] and *Caulobacter crescentus* CB15 [8]; all of which are important as pathogens or other research targets, and their genome sequences continue to be used as references, some of these citations are briefly described in Table 1. We used Illumina MiSeq 150 cycle, paired-end sequencing protocols to sequence their genomic DNAs obtained from ATCC (http://www.atcc.org). To obtain high sequence coverage for high CG% genomes, *Caulobacter crescentus* CB15 (67.2% GCs) and *Salmonella enterica* ATCC 9150 (52.2% GCs) were sequenced in a lane together and the other three (lower than 50% GCs) were sequenced in a separate lane together.

Results

Sequencing coverages were: 325X, 63X, 116X, 111X and 152X for *C.crescentus* CB15, *S. enterica* ATCC 9150, *B.fragilis* NCTC 9343, *B.halodurans* C-125 and *V.cholerae* O1 N16961, respectively (Table 2). Using BWA to map the sequence reads to the reference sequences of the corresponding genomes, we counted the number of loci covered by at

Table 1 Citations linked to originally sequenced bacterial genomes

Organism	Citations associated with originally published genome sequences
Salmonella enterica subsp. Enterica serovar Paratyphi A 9150	**227 Citations**
	Crump et al., 2010. Global trends in typhoid and paratyphoid Fever. Clin Infect Dis 50:241–246.
	Yang F, et al., 2005. Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. Nucleic acids research 33:6445–6458.
	Thomson NR, et., 2008. Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. Genome research 18:1624–1637.
V.cholerae O1 N16961	**1290 Citations**
	Thompson FL, et al., 2004. Biodiversity of vibrios. Microbiology and molecular biology reviews 68:403–431.
	Makino K, et al., 2003. Genome sequence of Vibrio parahaemolyticus: a pathogenic mechanism distinct from that of V cholerae. The Lancet 361:743–749.
	Zhu J, et al., 2002. Quorum-sensing regulators control virulence gene expression in Vibrio cholerae. Proceedings of the National Academy of Sciences 99:3129–3134.
	Merrell DS et., 2002. Host-induced epidemic spread of the cholera bacterium. Nature 417:642–645.
C. crescentus CB15	**417 Citations**
	Hu P, et al., 2005. Whole-genome transcriptional analysis of heavy metal stresses in Caulobacter crescentus. Journal of bacteriology 187:8437–8449.
	Laub MT, et al., 2002. Genes directly controlled by CtrA, a master regulator of the Caulobacter cell cycle. Proceedings of the National Academy of Sciences 99:44632–4637.
	Hottes AK, et al., 2005. DnaA coordinates replication initiation and cell cycle transcription in Caulobacter crescentus. Molecular microbiology 58:1340–1353.
	Reisenauer A, et al., 2002. DNA methylation affects the cell cycle transcription of the CtrA global regulator in Caulobacter. The EMBO journal 21:4969–4977.

Described are the number of total citations the original publication describing the sequenced genome was cited in and multiple select articles describing these data in related research.

least 10 reads of which at least half showed different read sequences from the references.

From the read alignments, we found 89, 17, 6, 147 and 165 loci of which read sequences were not consistent with the reference sequences for *C.crescentus* CB15, *S.enterica* ATCC 9150, *B.fragilis* NCTC 9343, *B.halodurans* C-125 and *V.cholerae* O1 N16961, respectively. All five reference sequences appeared to have loci covered by inconsistent read sequences, and the numbers of inconsistent loci were unexpectedly high for four bacteria, and modest for *B. fragilis* NCTC 9343. However, as we have shown in our previous studies of *Brucella* [9], not every inconsistent locus could be detected by the first alignment because alignment programs have limitations in properly aligning reads to loci containing repeat sequences, long INDELs or other structural differences. To detect structural assembly errors from read alignments, we inspected loci where at least 20% of the reads covering them were clipped (partially unaligned) at the same bases. About 4 ~ 20 loci covered by clipped reads were detected from read alignments of the five reference sequences. More than half of the loci were in the G/C homopolymer regions which frequently cause sequencing systems to generate incorrect

Table 2 Re-sequenced bacterial genomes from six organisms

Comparison of genomic sequencing quality between sanger & NGS methods

Organism	Genome size	First published year	Last updated year	Re-sequenced by illumina sequencer	
				Coverage	**Number of Inconsistent Loci**
Brucelli suis 1330	Chr.1 2.1 M Chr.2 1.2 M	2002	2011	1559X	12 loci
Caulobacter crescentus CB15	4.0 M	2001		325X	89 loci
Salmonella enterica ATCC 9150	4.6 M	2004		63X	17 loci
Bacteroides fragilis NCTC 9343	5.2 M	2002	2005	116X	6 loci
Bacillus halodurans C-125	4.2 M	2000	2005	111X	147 loci
Vibrio cholerae O1 N16961	Chr.1 2.9 M Chr.2 1.1 M	2000		152X	165 loci

The genome sequence of *Brucella suis* 1330 was re-sequenced using the Illumina GAIIx platform (previously published) and the additional five genomes were sequenced using Illumina MiSeq platform.

random sequences, thus unaligned parts of read sequences were not consistent. At other loci, the unaligned parts of read sequences were consistent and able to generate consensus sequences, which are duplications of other loci or do not exist in the reference sequences, indicating potential structural assembly errors (or they may be the results of rapid evolutionary changes).

Conclusions

The usage of genomic sequencing material derived from Sanger sequencing methods were a valuable, pioneering tool towards current methods. However, this method is highly error prone and the continued use of these sequenced genomes to identify anomalous and unique genomic traits could be additive in error to original findings, unless these sequences are updated. A few, such as *Escherichia coli* K-12 sub-strain MG1655, have been continuously updated by the original submitters, but many completed sequences contain assembly errors and lack necessary revisions. Species specific genome sequences are used in a variety of platforms in basic and applied research, including: understanding evolutionary relationships, mechanisms of microbial virulence and disease pathogenesis, diagnostics, and food and health safety. As a scientific community, we are able to illustrate the needs and the capability to rectify these errors by next-gen, re-sequencing as seen in the reanalysis of multiple organisms [1,2]. Now, with advances in NGS technologies which can generate tremendous amounts of raw sequencing data in a cost and time efficient way, high sequence coverage of bacterial genomes has been enabled to validate these data and revise single nucleotide or short INDEL errors. In this small study we have successfully demonstrated that these errors can be minimized with NGS methods and also propose a concerted initiative to re-sequence genomes from the 'Sanger-era'[9]. As concerns for reproducibility in science are ever increasing- with special emphasis linked to 'big-data' and genomics- science must address sequenced microbial genomes and establish standards for highlighting older sequenced material and flagging these data to be used with caution. This is a contemporary issue, for the genomes previously measured years ago are still very much in use, a current solution (and investment to science) is nextgen re-sequencing. By conducting large scale evaluations of genome sequences published during the early 2000s, as a scientific community we would safeguard public interests and the integrity of future endeavors from the consequence of existing errors.

Competing interests

These data have not been previously published and are not being considered for publication elsewhere. Dr. Harold R. Garner is a co-owner of Genomeon, L.L.C., a startup company which may enter into an exclusive licensing agreement with Virginia Tech for these data. Genomeon played no role or influenced the direction of this research in any way.

Authors' contributions

HT is a software-developer and programmer who conducted the genomic analyses described and identified sequencing errors importantly discussed in the manuscript. EK and JHB contributed equally to biological interpretations, microbe selections for analysis, and manuscript development. HRG contributed significantly to the design and intellectual development of the study. All authors read and approved the final manuscript.

Acknowledgements

This work was funded by the Medical Informatics and Systems Division directors fund to Dr. Garner at VBI/Virginia Tech. We thank the system administrators in the VBI computational core (Jason Decker, Michael Snow, Dominik Borkowski, David Bynum, Douglas McMaster, Jeremy Johnson, and Vedavyas Duggirala) for technical support.

Received: 11 July 2014 Accepted: 25 October 2014
Published: 7 November 2014

References

1. Tae H, Shallom S, Settlage R, Preston D, Adams LG, Garner HR: **Revised genome sequence of Brucella suis 1330.** *J Bacteriol* 2011, **193:**6410.
2. Paulsen IT, Seshadri R, Nelson KE, Eisen JA, Heidelberg JF, Read TD, Dodson RJ, Umayam L, Brinkac LM, Beanan MJ, Daugherty SC, Deboy RT, Durkin AS, Kolonay JF, Madupu R, Nelson WC, Ayodeji B, Kraul M, Shetty J, Malek J, Van Aken SE, Riedmuller S, Tettelin H, Gill SR, White O, Salzberg SL, Hoover DL, Lindler LE, Halling SM, Boyle SM, Fraser CM: **The Brucella suis genome reveals fundamental similarities between animal and plant pathogens and symbionts.** *Proc Natl Acad Sci U S A* 2002, **99:**13148–13153.
3. Wynne JW, Seemann T, Bulach DM, Coutts SA, Talaat AM, Michalski WP: **Resequencing the Mycobacterium avium subsp. paratuberculosis K10 Genome: Improved Annotation and Revised Genome Sequence.** *J Bacteriol* 2010, **192:**6319–6320.
4. Cerdeno-Tarraga AM, Patrick S, Crossman LC, Blakely G, Abratt V, Lennard N, Poxton I, Duerden B, Harris B, Quail MA, Barron A, Clark L, Corton C, Doggett J, Holden MT, Larke N, Line A, Lord A, Norbertczak H, Ormond D, Price C, Rabbinowitsch E, Woodward J, Barrell B, Parkhill J: **Extensive DNA inversions in the B. fragilis genome control variable gene expression.** *Science* 2005, **307:**1463–1465.
5. McClelland M, Sanderson KE, Clifton SW, Latreille P, Porwollik S, Sabo A, Meyer R, Bieri T, Ozersky P, McLellan M, Harkins CR, Wang C, Nguyen C, Berghoff A, Elliott G, Kohlberg S, Strong C, Du F, Carter J, Kremizki C, Layman D, Leonard S, Sun H, Fulton L, Nash W, Miner T, Minx P, Delehaunty K, Fronick C, Magrini V, Nhan M, Warren W, Florea L, Spieth J, Wilson RK: **Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of Salmonella enterica that cause typhoid.** *Nat Genet* 2004, **36:**1268–1274.
6. Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Umayam L, Gill SR, Nelson KE, Read TD, Tettelin H, Richardson D, Ermolaeva MD, Vamathevan J, Bass S, Qin H, Dragoi I, Sellers P, McDonald L, Utterback T, Fleishmann RD, Nierman WC, White O, Salzberg SL, Smith HO, Colwell RR, Mekalanos JJ, Venter JC, Fraser CM: **DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae.** *Nature* 2000, **406:**477–483.
7. Takami H, Nakasone K, Takaki Y, Maeno G, Sasaki R, Masui N, Fuji F, Hirama C, Nakamura Y, Ogasawara N, Kuhara S, Horikoshi K: **Complete genome sequence of the alkaliphilic bacterium Bacillus halodurans and genomic sequence comparison with Bacillus subtilis.** *Nucleic Acids Res* 2000, **28:**4317–4331.
8. Nierman WC, Feldblyum TV, Laub MT, Paulsen IT, Nelson KE, Eisen JA, Heidelberg JF, Alley MR, Ohta N, Maddock JR, Potocka I, Nelson WC, Newton A, Stephens C, Phadke ND, Ely B, DeBoy RT, Dodson RJ, Durkin AS, Gwinn ML, Haft DH, Kolonay JF, Smit J, Craven MB, Khouri H, Shetty J, Berry K, Utterback T, Tran K, Wolf A, Vamathevan J, Ermolaeva M, White O, Salzberg SL, Venter JC, Shapiro L, Fraser CM: **Complete genome sequence of Caulobacter crescentus.** *Proc Natl Acad Sci U S A* 2001, **98:**4136–4141.
9. Tae H, Settlage RE, Shallom S, Bavarva JH, Preston D, Hawkins GN, Adams LG, Garner HR: **Improved variation calling via an iterative backbone remapping and local assembly method for bacterial genomes.** *Genomics* 2012, **100:**271–276.

doi:10.1186/1756-0381-7-25
Cite this article as: Tae *et al.*: Updating microbial genomic sequences: improving accuracy & innovation. *BioData Mining* 2014 **7**:25.