JOURNAL OF CLINICAL MICROBIOLOGY, Jan. 2011, p. 23–26 Vol. 49, No. 1
0095-1137/11/$12.00 doi:10.1128/JCM.01918-10

COMMENTARY

Call for a Quality Standard for Sequence-Based Assays in Clinical Microbiology: Necessity for Quality Assessment of Sequences Used in Microbial Identification and Typing[▽]

Anthony Underwood* and Jonathan Green

Bioinformatics Group, Microbiological Services Division, Health Protection Agency, London, United Kingdom

Subsequent to the early days of radioisotope sequencing, DNA sequencing technologies have evolved rapidly, giving rise to techniques that are more sensitive and rapid and provide longer sequence reads. These incremental improvements in technology and in the automation of modern sequencing platforms have resulted in applications of DNA sequencing across virtually all biological science disciplines. In microbiology, this has led to a significant increase in the number of available complete microbial and viral genome sequences and, consequently, to increasing use of DNA sequence analysis in clinical microbiology laboratories, where the results may be influential on consequent patient management. The evolution of Sanger-based technologies is now being complemented by the revolution of second- and third-generation sequencing technologies that, potentially, could provide whole-microbial-genome analysis at the clinical bench and have a greater, more direct impact on patient and outbreak management. While the technologies have continued to evolve, common standards for the description of DNA sequence data quality still do not exist. Further, utilization of DNA sequence data requires comparison with other similar sequence data held in private or public databases. Similarly, there is no process for the accreditation of these databases and the analyses that they provide to ensure that they are fit for purpose. We strongly urge that action is required to address this.

CURRENT USES OF SEQUENCING IN CLINICAL MICROBIOLOGY

DNA sequencing is being introduced into microbial diagnosis particularly for the identification of esoteric organisms where traditional methods fail or where the delay to definitive identification by traditional means is significant to patient treatment. The most common locus used for bacterial identification is the 16S rRNA gene (6). In general, the complete sequence data rather than individual bases are important when making comparisons to existing data to make the identification. Other loci, in particular *rpoB* and *gyrB* (19, 31), are often used for genera whose species are more difficult to resolve with the 16S locus. Because the identification may be used in clinical management to alter treatment regimens, the quality of the data used to make this identification is crucial.

Sequence-based typing methodologies, typified by multilocus sequencing typing (MLST), are often used for typing in epidemiological studies and population biology. The results do not usually have a clinical impact in the short term but may be of great importance in medico-legal cases when identifying source of infection, and therefore data quality is essential. Crucially, with these techniques, a single-base change alters the type; therefore, the quality at individual bases is as relevant as overall quality.

Sequencing has not been widely used for determining antibiotic resistance in bacteria, because often the presence or absence of a gene or genes is sufficient to convey resistance, or regulator genes make it more than a single gene trait. Because bacteria are typically easily and quickly cultured, phenotypic testing usually produces an inexpensive, definitive result, making molecular methods, including sequencing, inappropriate (32). The exception is for slow-growing bacteria, such as *Mycobacterium tuberculosis*, where a phenotypic result may take several weeks to produce. *M. tuberculosis* is also of particular relevance because it does not acquire DNA, including resistance genes, horizontally, and so all resistance is the result of chromosomal mutations. In this case, sequencing of the genes involved in rifampin and isoniazid resistance may yield a result within hours (11). When sequencing these loci, the high quality of the bases causing potential resistance is obviously essential, since an erroneous base will incorrectly inform the clinician of the resistance profile of the organism causing the infection and therefore may result in inappropriate treatment.

In contrast to bacteria, where sequencing has only recently become more prevalent for the purposes of identification and typing, sequencing has been the gold standard for some time for viral typing and to a lesser extent viral identification. For hepatitis C virus (HCV), genotyping has shown greater ability to discriminate between isolates than other methods. Another important example is HIV, where subtyping has revealed several important findings with significance for surveillance, transmission studies, and vaccine design (27). Genotypic analysis of HIV resistance is the most common means used to predict HIV resistance but is best used in conjunction with phenotypic testing (30).

* Corresponding author. Mailing address: Bioinformatics Group, Microbiological Services Division (Colindale), Health Protection Agency, 61 Colindale Avenue, London NW9 5EQ, United Kingdom. Phone: 44 208 3276466. Fax: 44 208 2051630. E-mail: anthony.underwood@hpa.org.uk.

[▽] Published ahead of print on 10 November 2010.

DEFINING DNA SEQUENCE QUALITY

In order to develop a common language describing sequence quality and thereby establish a common standard, we need to examine the process of base calling, the process of converting electrophoretic signals into a DNA sequence.

Originally, most sequencing was performed using a radioactive-based Sanger dideoxy sequencing technique. For reasons of safety, ease, and automation, this has largely been replaced by fluorescent dye-based Sanger dideoxy sequencing, where the products are run either on acrylamide gels or capillaries containing a viscous polymer. The output from this technique is a fluorescent trace file that can be processed to produce a final output which is a string of alphanumeric characters, each of which will be one of the four nucleotide bases (G, A, T, or C) if none of the bases are ambiguous. Although these endpoint data are easily comparable between labs, the process of converting the fluorescent trace files to the alphanumeric string is performed by an algorithm on the sequencing machine in a process named base calling. In essence, each of the algorithms is designed to be able to identify real peaks based on the features of the trace. In areas where the sequence is of poor quality, the algorithm will either not be able to assign a base or may assign an incorrect base. This problem can sometimes be rectified by manual editing of the data after computer base calling. Many scientists consider manual assessment of sequence to be the gold standard when applied to the process of ensuring optimum sequence quality (20). However, different scientists may use different subjective standards when deciding a sequence to be of adequate quality for the desired purpose. A recent study on the need for quality assurance in DNA sequencing (1) concluded that, due to the large number of data points that need to be analyzed in DNA sequencing, there is an increased risk of error and that this emphasizes the need for automatic quality-assessed base calling of DNA sequence and automated genotype/mutation assessment.

Automated quality assessment, in contrast to manual approaches, allows quantitative values for quality to be assigned to any sequence. Trimming and/or acceptance/rejection of a particular DNA sequence can therefore be based on numerical cutoff values, and this permits straightforward comparison of results between laboratories. Phred is considered to be the gold-standard software for automated sequence quality assessment. Phred reads DNA sequencing trace files, calls bases, and assigns a quality value to each called base. The Phred quality values have been thoroughly tested for both accuracy and power to discriminate between correct and incorrect base calls (7, 8). Initial testing achieved 40 to 50% lower error rates on large test data sets than other software (8). The assignment of error probabilities allows for quantitative benchmarking of different sequencing methods and protocol changes. Phred was designed to be incorporated in "pipelines" in the sequencing workflow of large sequencing centers and is widely used by the larger academic and commercial sequencing laboratories. To date, few desktop applications have used the Phred algorithm, and this makes it inaccessible to the average user. A wide range of software is available to fulfil a similar purpose, and each may have its own metric for the description of quality, making comparisons between software difficult or impossible. Sequencing platforms from both Beckman and ABI have associated software that have improved dramatically compared to an initial comparison with Phred and can now assign base quality scores on the same scale as Phred, with the facility for trimming sequences based on quality.

Sequencher is a piece of software often used in clinical laboratories because of its ability to quickly analyze a set of sequences for single nucleotide polymorphisms (SNPs) and to take account of quality scores derived from base calling by its own built-in algorithm or by Phred when making judgments about SNPs.

Forthcoming changes to the *in vitro* diagnostic medical devices directive (IVDD) are likely to include standalone software within its scope (http://www.edma-ivd.be/fileadmin/upl_documents/Position_Papers/software_29jan04.doc). This means that bioinformatics software will almost certainly have to be CE marked, and assessment of sequence quality will surely be part of this.

DATA FROM NEW TECHNOLOGIES

A number of new sequencing technologies have arisen that have already been demonstrated to be useful in public health microbiology. These include pyrosequencing (22–24), a technique with current clinical applications (2, 4, 15, 28) producing short reads for accurate and quantitative analysis of DNA sequences, including built-in proprietary quality scores, and Sequenom, a proprietary technology using the MassARRAY system to perform reliable SNP analysis. Sequenom is said to be 99.7% accurate and provides quality values for each assay, but these are not comparable with other systems due to the different nature of the technology. The technique is inexpensive and has an extremely high throughput. Initial results where this technology has been used for bacterial typing show promise (14).

"Next-generation sequencing" refers to a group of technologies that currently allow rapid sequencing of DNA templates (often whole genome) through massively parallel reactions (17, 25, 26). Because of the different data types generated, it will prove difficult to design a single common nomenclature for quality. Instead standards will arise for each platform (3, 5). Deriving clinically useful information from this sequence will present many bioinformatics challenges, not least of which will be to assess the quality that can be assigned to any result; however, there are already some examples of the utility of this type of data in a clinical setting (13).

USE OF EXTERNAL DATABASES WHEN ANALYZING RESULTS

Use of DNA sequencing in a clinical context, whether for diagnostics, typing, or antimicrobial/viral resistance analysis, always requires comparison with another data source of reference sequences, usually a local or online database. The validity of the performed comparative analysis is not only determined by the quality of the submitted DNA sequence but also by the quality of the sequences with which it is compared. Although the former can be quality controlled (QC) by the local scientists, the latter is dependent on the policy of those who curate the databases.

Some databases, including those of the International Sequence Database Collaboration (a partnership of three centers comprising the largest annotated collection of all publicly available DNA sequences), include little or no quality control on the data that they contain. It is stated on the GenBank website that "GenBank depends on its contributors to help keep the database as comprehensive, current, and accurate as possible." This policy has led to a significant number of sequences being misannotated and/or of poor quality, containing multiple ambiguous bases. This means that, although GenBank represents an excellent resource for research, comparison of laboratory results with data from GenBank in a clinical setting may result in inaccurate and potentially harmful results.

There are some databases that have attempted to address quality issues by applying quality control procedures which may be manual and/or automated. For example, The Stanford HIV drug resistance database, designed to represent, store, and analyze the divergent forms of data underlying HIV drug resistance (21), provides resources to correlate genotype to response to treatment, genotype to drug resistance phenotype, and genotype to clinical outcome. The database is curated by manual processes to ensure unambiguous sequence data and accurate annotation of the sequence in relation to clinical data. Another curated database is the Influenza Research Database (IRD) that superseded the Influenza Sequence Database held at Los Alamos National Laboratory (16). During the curation process, care is taken to remove redundant sequences and to fill in missing field data. MLST is used widely for typing of medically important bacterial strains (www.mlst.net and pubmlst.org). The raw trace files of sequences for new alleles from each organism are subject to manual curation. Because this is performed by different individuals, perhaps with different approaches and standards that they apply, it will be difficult to ensure a consistent metric for sequence quality across databases.

Some microbial databases utilize an automated curation pipeline before secondary manual checking. The RIDOM database for 16S rRNA genes uses a combination of FASTA, CLUSTAL W, and a Phred/Phrap pipeline to produce a comprehensive, high-quality database of 16S sequences for medically important organisms (12) and has been shown, for the identification of *Nocardia* at least, to outperform other databases for both specificity and sensitivity (18). An online tool developed for the *Legionella pneumophila* sequence-based typing scheme (9, 10) also utilizes an automated curation filter which allows submission of trace files and automated acceptance or rejection of these sequences based on Phred-based quality score cutoffs. The application allows for a more streamlined quantitative curation of the data submitted and gives a transparent, quantitative metric for the data within the publicly accessible database (29).

PROPOSAL

To date, the situations where utilization of DNA sequencing of microbes directly impact clinical management of patients are small in number but individually significant. Molecular typing by DNA sequencing is already important for surveillance, and the result may influence how outbreaks are managed. It is surprising therefore that DNA sequence data quality standards are not better defined and that there are not well-defined guidelines about best practices for downstream analysis procedures in these situations. It is anticipated that next-generation or, more likely, third-generation sequencing technologies have the potential to provide routine whole-genome identification and typing (and perhaps antimicrobial resistance typing) within the next few years. It is essential that, for these emerging sequencing technologies, the scientific community ensures that producing consistent and comparable quality scores is not neglected among the many other challenges that they pose.

In addition to a common metric or language for DNA sequence quality, we need also to consider reference databases that contribute to the interpretation of these data, including the quality of the data that they hold, the analytical rigor that they provide, and the ease of interpretation of results. Defining these for a particular purpose will be a sensible step toward ensuring that the outcome of the laboratory investigation best serves the patient.

Written procedures alone are probably not enough, and there will be a significant role for external quality assurance (EQA) in ensuring competence. EQUALseq is a European Union-funded initiative whose aim is to develop methodological EQA schemes for sequencing (1). One of its most striking results was the diversity in sequencing performance between laboratories. The results demonstrated that those performing over 1,000 sequencing assays per year produced significantly better results, emphasizing the importance of the technical skill and experience in determining high-quality output. Not only was the technical expertise in performing the sequencing reaction itself crucial, but the need for expert postanalytical proficiency was also highlighted. Another European program, European Molecular Genetics Quality Network (EQMN), was funded by the United Kingdom Department of Health to develop an EQA scheme to evaluate DNA sequence analysis and surveyed the quality of DNA sequencing from 64 laboratories from 21 countries (20). In the majority of cases, the sequence data generated exceeded the recommended Phred score of 20 (99% confidence). Often the mistakes in genotyping did not correlate with the quality of the sequence but rather misassignation of genotype due to human error. These studies suggest that with the increasing importance of sequencing in clinical laboratory diagnostics, evidence of competency for those providing this service should be obligatory. One means of providing this evidence is accreditation to a national (for example, Clinical Pathology Accreditation in the United Kingdom) or international (ISO 15189 or ISO 17025) standard, both of which require participation in EQA. These studies also emphasize the need for automatic quality-assessed base calling of DNA sequence and automated genotype/mutation assessment. Crucial for the efficacy of these schemes is quality-controlled test materials. The QC materials could either be microbial cultures to test for the whole procedure, from DNA extraction through PCR to sequencing, or pregenerated PCR products/synthetic DNA, just testing the sequencing methodology. Although EQA schemes will be essential, we also suggest that if the sequence used to generate the test result itself was subject to quality control on a per-test basis rather than per-laboratory basis, then many incorrect results could be avoided.

There are many factors to consider when developing DNA sequence quality standards that are flexible enough to cover the diverse range of tests used in clinical microbiology and also be acceptable across the community. We recommend that a working group be established, bringing together molecular and clinical microbiologists, quality experts, and commercial suppliers who could together propose a consensus set of standards for DNA sequence quality and a process for the implementation in clinical testing. These should include recommendations for the establishment of EQA schemes with appropriate QC materials. Without these standards, the trustworthiness of sequence-based tests may be called into question.

ACKNOWLEDGMENTS

We acknowledge Vivienne James and Ian Sharp for their helpful comments and suggestions on the contents of this article.

REFERENCES

1. **Ahmad-Nejad, P., A. Dorn-Beineke, U. Pfeiffer, J. Brade, W. J. Geilenkeuser, S. Ramsden, M. Pazzagli, and M. Neumaier.** 2006. Methodologic European external quality assurance for DNA sequencing: the EQUALseq program. Clin. Chem. **52:**716–727.
2. **Arnold, C., L. Westland, G. Mowat, A. Underwood, J. Magee, and S. Gharbia.** 2005. Single-nucleotide polymorphism-based differentiation and drug resistance detection in Mycobacterium tuberculosis from isolates or directly from sputum. Clin. Microbiol. Infect. **11:**122–130.
3. **Bravo, H. C., and R. A. Irizarry.** 2009. Model-based quality assessment and base-calling for second-generation sequencing data. Biometrics **66:**665–674.
4. **Bright, R. A., D. K. Shay, B. Shu, N. J. Cox, and A. I. Klimov.** 2006. Adamantane resistance among influenza A viruses isolated early during the 2005-2006 influenza season in the United States. JAMA **295:**891–894.
5. **Brockman, W., P. Alvarez, S. Young, M. Garber, G. Giannoukos, W. L. Lee, C. Russ, E. S. Lander, C. Nusbaum, and D. B. Jaffe.** 2008. Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res. **18:**763–770.
6. **Clarridge, J. E., III.** 2004. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin. Microbiol. Rev. **17:**840–862.
7. **Ewing, B., and P. Green.** 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8:**186–194.
8. **Ewing, B., L. Hillier, M. C. Wendl, and P. Green.** 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:**175–185.
9. **Gaia, V., N. K. Fry, B. Afshar, P. C. Luck, H. Meugnier, J. Etienne, R. Peduzzi, and T. G. Harrison.** 2005. Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of Legionella pneumophila. J. Clin. Microbiol. **43:**2047–2052.
10. **Gaia, V., N. K. Fry, T. G. Harrison, and R. Peduzzi.** 2003. Sequence-based typing of Legionella pneumophila serogroup 1 offers the potential for true portability in legionellosis outbreak investigation. J. Clin. Microbiol. **41:** 2932–2939.
11. **Garcia de Viedma, D.** 2003. Rapid detection of resistance in Mycobacterium tuberculosis: a review discussing molecular approaches. Clin. Microbiol. Infect. **9:**349–359.
12. **Harmsen, D., J. Rothganger, M. Frosch, and J. Albert.** 2002. RIDOM: Ribosomal Differentiation of Medical Micro-organisms Database. Nucleic Acids Res. **30:**416–417.
13. **Harris, S. R., E. J. Feil, M. T. G. Holden, M. A. Quail, E. K. Nickerson, N. Chantratita, S. Gardete, A. Tavares, N. Day, J. A. Lindsay, J. D. Edgeworth, H. de Lencastre, J. Parkhill, S. J. Peacock, and S. D. Bentley.** 2010. Evolution of MRSA during hospital transmission and intercontinental spread. Science **327:**469–474.
14. **Honisch, C., Y. Chen, C. Mortimer, C. Arnold, O. Schmidt, D. van den Boom, C. R. Cantor, H. N. Shah, and S. E. Gharbia.** 2007. Automated comparative sequence analysis by base-specific cleavage and mass spectrometry for nucleic acid-based microbial typing. Proc. Natl. Acad. Sci. U. S. A. **104:**10649–10654.
15. **Jordan, J. A., A. R. Butchko, and M. B. Durso.** 2005. Use of pyrosequencing of 16S rRNA fragments to differentiate between bacteria responsible for neonatal sepsis. J. Mol. Diagn. **7:**105–110.
16. **Macken, C., H. Lu, J. Goodman, and L. Boykin.** 2001. The value of a database in surveillance and vaccine selection, p. 103–106. *In* A. D. M. E. Osterhaus, N. Cox, and A. W. Hampson (ed.), Options for the control of influenza IV. Elsevier Science, Amsterdam, Netherlands.
17. **Mardis, E. R.** 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. **24:**133–141.
18. **Mellmann, A., J. L. Cloud, S. Andrees, K. Blackwood, K. C. Carroll, A. Kabani, A. Roth, and D. Harmsen.** 2003. Evaluation of RIDOM, MicroSeq, and Genbank services in the molecular identification of Nocardia species. Int. J. Med. Microbiol. **293:**359–370.
19. **Mollet, C., M. Drancourt, and D. Raoult.** 1997. *rpoB* sequence analysis as a novel basis for bacterial identification. Mol. Microbiol. **26:**1005–1011.
20. **Patton, S. J., A. J. Wallace, and R. Elles.** 2006. Benchmark for evaluating the quality of DNA sequencing: proposal from an international external quality assessment scheme. Clin. Chem. **52:**728–736.
21. **Rhee, S. Y., M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer.** 2003. Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res. **31:**298–303.
22. **Ronaghi, M.** 2003. Pyrosequencing for SNP genotyping. Methods Mol. Biol. **212:**189–195.
23. **Ronaghi, M.** 2001. Pyrosequencing sheds light on DNA sequencing. Genome Res. **11:**3–11.
24. **Ronaghi, M., and E. Elahi.** 2002. Pyrosequencing for microbial typing. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. **782:**67–72.
25. **Schuster, S. C.** 2008. Next-generation sequencing transforms today's biology. Nat. Methods **5:**16–18.
26. **Shendure, J., and H. Ji.** 2008. Next-generation DNA sequencing. Nat. Biotechnol. **26:**1135–1145.
27. **Tatt, I. D., K. L. Barlow, A. Nicoll, and J. P. Clewley.** 2001. The public health significance of HIV-1 subtypes. AIDS **15**(Suppl. 5)**:**S59–S71.
28. **Tuohy, M. J., G. S. Hall, M. Sholtis, and G. W. Procop.** 2005. Pyrosequencing as a tool for the identification of common isolates of Mycobacterium sp. Diagn. Microbiol. Infect. Dis. **51:**245–250.
29. **Underwood, A., W. Bellamy, B. Afshar, and N. Fry.** 2006. Development of an online tool for the European Working Group for Legionella Infections sequence-based typing, including automatic quality assessment and data submission, p. 163–166. *In* N. Cianciotto, Y. Abu Kwaik, P. Edelstein, B. Fields, D. Geary, T. Harrison, C. Joseph, R. Ratcliff, J. Stout, and M. Swanson (ed.), *Legionella*: state of the art 30 years after its recognition. ASM Press, Washington, DC.
30. **Van Laethem, K., and A. M. Vandamme.** 2006. Interpreting resistance data for HIV-1 therapy management—know the limitations. AIDS Rev. **8:**37–43.
31. **Watanabe, K., J. Nelson, S. Harayama, and H. Kasai.** 2001. ICB database: the *gyrB* database for identification and classification of bacteria. Nucleic Acids Res. **29:**344–345.
32. **Woodford, N., and A. Sundsfjord.** 2005. Molecular detection of antibiotic resistance: when and where. J. Antimicrob. Chemother. **56:**259–261.

The views expressed in this Commentary do not necessarily reflect the views of the journal or of ASM.