A Joint Announcement on Genome Sequence Standards

Genome project standards in a new era of sequencing

P. S. G. Chain[1,2,3,4,22,*], D. V. Grafham[5,*], R. S. Fulton[6], M. G. FitzGerald[7], J. Hostetler[8], D. Muzny[9], J. C. Detter[1,10], J. Ali[11], B.Birren[7], D. C. Bruce[1, 10], C. Buhay[9], J. R. Cole[3,4], Y. Ding[9], S. Dugan[9], D. Field[12], G. M. Garrity[3,4], R. Gibbs[9], T. Graves[6], C. S. Han[1, 10], S. H. Harrison[3], S. Highlander[9], P. Hugenholtz[1], H. M. Khouri[13], C. D. Kodira[7,23], E. Kolker[14,15], N. C. Kyrpides[1], D. Lang[1,2], A. Lapidus[1], S. A. Malfatti[1,2], V. Markowitz[16], T. Metha[7], K. E. Nelson[8], J. Parkhill[5], S. Pitluck[1], X. Qin[9], T. D. Read[17], J. Schmutz[18], S. Sozhamannan[19], R. Strausberg[8], G. Sutton[8], N. R. Thomson[5], J. M. Tiedje[3,4], G. Weinstock[6], A. Wollam[6], and the entire GSC[20] and HMP Jumpstart[21] consortia.

[1]U.S. Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA
[2]Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, Livermore, California 94550, USA
[3]Microbiology & Molecular Genetics, Michigan State University, East Lansing, Michigan 48824, USA
[4]Center for Microbial Ecology, Michigan State University, East Lansing, Michigan 48824, USA
[5]The Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom
[6]The Genome Center, Washington University School of Medicine, St Louis, Missouri 63108, USA
[7]The Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts 02141, USA
[8]J. Craig Venter Institute, Rockville, Maryland 20850, USA
[9]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA
[10]Bioscience Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA
[11]Ontario Institute for Cancer Research, Toronto, Ontario M5G 0A3, Canada
[12]Natural Environmental Research Council Centre for Ecology and Hydrology, Oxford, Oxfordshire OX1 3SR, UK
[13]National Center for Biotechnology Information, National Library of Medicine, Rockville, Maryland 20850, USA
[14]Seattle Children's Hospital and Research Institute, Seattle, Washington 98101, USA
[15]Biomedical & Health Informatics Division, MEBI, University of Washington School of Medicine, Seattle, Washington 98195, USA
[16]Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
[17]Emory GRA Genomics Core, Emory University School of Medicine, Atlanta, Georgia 30322, USA
[18]HudsonAlpha Genome Sequencing Center, HudsonAlpha Institute, Huntsville, Alabama 35806, USA
[19]Biological Defense Research Directorate, Naval Medical Research Center, Silver Spring, Maryland 20910, USA
[20] Genomic Standards Consortium
[21] Human Microbiome Project Jumpstart Consortium
[22] Current address: Bioscience Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA
[23] Current address: 454 Life Sciences, Branford, Connecticut 06405, USA

*Address correspondence to Patrick Chain (pchain@lanl.gov) and Darren Grafham (dg1@sanger.ac.uk)

For over a decade, genome sequences have adhered to only two standards that are relied on for purposes of sequence analysis by interested third parties (1, 2). However, ongoing developments in revolutionary sequencing technologies have resulted in a redefinition of traditional whole genome sequencing that requires a careful reevaluation of such standards. With commercially available 454 pyrosequencing (followed by Illumina, SOLiD, and now Helicos), there has been an explosion of genomes sequenced under the moniker 'draft', however these can be very poor quality genomes (due to inherent errors in the sequencing technologies, and the inability of assembly programs to fully address these errors). Further, one can only infer that such draft genomes may be of poor quality by navigating through the databases to find the number and type of reads deposited in sequence trace repositories (and not all genomes have this available), or to identify the number of contigs or genome fragments deposited to the database. The difficulty in assessing the quality of such deposited genomes has created some havoc for genome analysis pipelines and contributed to many wasted hours of (mis)interpretation.

These same novel sequencing technologies have also brought an exponential leap in raw sequencing capability, and at greatly reduced prices that have further skewed the time- and cost-ratios of draft data generation versus the painstaking process of improving and finishing a genome. The resulting effect is an ever-widening gap between drafted and finished genomes that only promises to continue (Figure 1), hence there is an urgent need to distinguish good and poor datasets. The sequencing institutes in the authorship, along with the NIH's Human Microbiome Project Jumpstart Consortium (3), strongly believe that a new set of standards is required for genome sequences. The following represents a set of six community-defined categories of genome sequence standards that better reflect the quality of the genome sequence, based on our collective understanding of the different technologies, available assemblers, and the varied efforts to improve upon drafted genomes. Due to the increasingly rapid pace of genomics we avoided the use of rigid numerical thresholds in

our definitions to take into account the types of products achieved by any combination of technology, chemistry, assembler, or improvement/finishing process.

Standard Draft: minimally or unfiltered data from any number of different sequencing platforms, that are assembled into contigs. This is the minimum standard for a submission to the public databases. Sequence of this quality will likely harbor many regions of poor quality and can be relatively incomplete. It may not always be possible to remove contaminating sequence data. Despite its shortcomings, Standard Draft is the least expensive to produce and still possesses useful information.

High Quality Draft: overall coverage representing at least 90% of the genome or target region. Efforts should be made to include only sequence of the target organism and exclude contaminating sequences. This is still a draft assembly with little or no manual review of the product. Sequence errors and misassemblies are possible, with no implied order and orientation to contigs. This level is appropriate for general assessment of gene content.

Improved High Quality Draft: additional work has been performed beyond the initial shotgun sequencing and High Quality Draft assembly, by using either manual or automated methods. This standard should contain no discernable misassemblies, and should have undergone some form of gap resolution to reduce the number of contigs and supercontigs (or scaffolds). Undetectable misassemblies are still possible, particularly in repetitive regions. Low quality regions and potential base errors may also be present. This product is normally adequate for comparison to other genomes.

Annotation-directed Improvement: may overlap with the previous standards, but the term emphasizes the verification and correction of anomalies within coding regions such as frameshifts, and stop codons. This standard will most often be used in cases involving complex genomes where improvement beyond this category fails to outweigh the associated costs. Gene models (gene calls, including intron/exon determination for eukaryotes) and annotation of the genomic content should fully support the biology of the organism and the scientific questions being investigated. Exceptions to this gene-

specific finishing standard should be noted with comments in the submission. Repeat regions at this level are not resolved, so errors in those regions are much more likely. This standard is useful for gene comparisons, alternative splicing analysis, and pathway reconstruction.

Non-contiguous Finished: describes high quality assemblies that have been subject to automated and manual improvement, and where closure approaches have been successful for almost all gaps, as well as misassembled and low quality regions, however some exceptions exist. All gaps and sequence uncertainties have been attempted to be resolved, and only those recalcitrant to resolution remain, but are specifically noted in the genome submission as to the nature of the uncertainty. This product is thus of Finished quality with the only exception being repetitive or intractable gaps, along with heterochromatic sequence for eukaryotic applications, thus making it appropriate for most analyses. For nearly all higher organisms, this is actually the grade that was previously called "Finished."

Finished: refers to the current gold standard; genome sequences with less than 1 error per 100,000 bp and where each replicon is assembled into a single contiguous sequence with a minimal number of possible exceptions commented in the submission record. All sequences are complete and have been reviewed and edited, all known misassemblies have been resolved, and repetitive sequences have been ordered and correctly assembled. Any remaining exceptions to highly accurate sequence within the euchromatin are commented in the submission. The Finished product is appropriate for all types of detailed analyses and acts as a high quality reference genome for comparative purposes. Some microbial genome sequences where multiple platforms have been used for the same genome have exceeded this standard, and it is believed that no bases are incorrect except for natural low-level biological variation.

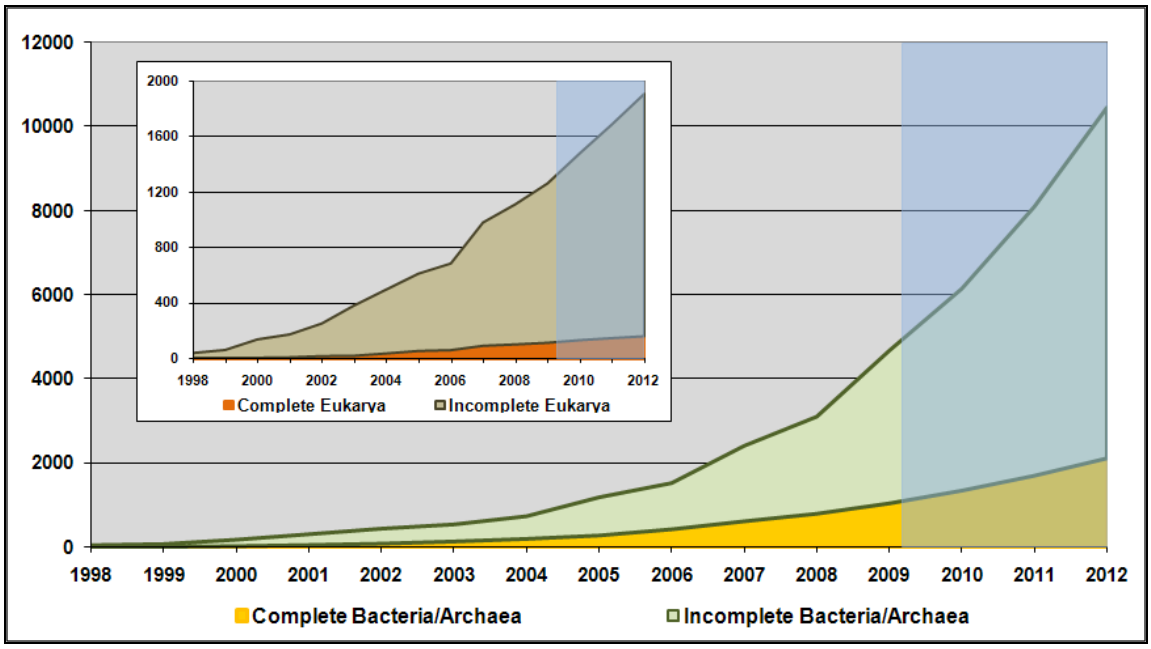
Intermediate standards often overlap, and while we do not advocate any one standard over another, we recommend that the target standard be based on the needs and goals of each project. There may be cases where select regions will be targeted for improvement and thus more than one of these standards may apply (such regionally

improved sequences should be identified within genome entries). This approach is most often used for eukaryotic whole genome sequencing projects, where the cost of complete finishing remains prohibitive and allows improvement to be directed at euchromatic sequence, since heterochromatic sequence remains largely recalcitrant to available approaches. Legacy eukaryotic tiling path standards will remain in use for a time to allow completion of some key projects.

Here, we have attempted to capture in a technology-independent fashion the types of whole genome sequencing projects that are beginning to populate databases and we have defined a set of standards that accommodate a growing list of alternative genome products that have been obtained via less conventional means, such as environmental (metagenomic) or single cell sequencing. Ongoing discussions with genome database repositories have been met with enthusiasm and the implementation of these standards as a requirement for genome submissions is expected. To aid in adoption of this classification of sequence finishing standards, we have added this classification to the Sequence Ontology where it can now be used to comply with the Genomic Standards Consortium's (GSC) "Minimum Information about a Genome Sequencing", or MIGS, standard (4) "sequencing status" descriptor. Furthermore, the efforts described here have recently been adopted under the umbrella of the GSC to ensure these efforts dovetail with related standardization efforts in the domain of genomics (5). This common currency in defining the products of genome projects enables better management of expectations in the research community and allows users of genomic data to assess the quality of the deposited available sequences and decide whether these meet their needs.

Figure Legend

Figure 1. Trends in generation of incomplete and complete genomes. Trends for bacterial and archaeal projects as well as those for eukaryotic projects (inset) are shown (a conservative estimate of future projects is projected following the trends of the past few months - shaded light blue). The data were derived from www.genomesonline.org (6).



References and Notes

1. International standards for sequence fidelity, established at the Second International Strategy Meeting on Human Genome Sequencing in Bermuda in 1997. “Finished” quality standards, commonly known as the Bermuda standards, defined finished sequence as a contiguous sequence with less than one error per 10,000 bases. Almost everything else was “draft”. Although intended for clone-based human genome sequencing, this finishing standard has been adopted for virtually all other genome projects. A number of exceptions and rules were later agreed upon (http://www.genome.gov/10001812), and these have undergone continuous refinement as Sanger sequencing technology matured but until recently have not required a significant re-evaluation.
2. Blakesley, R.W. et al. *Genome Res* **14**, 2235 (2004).
3. http://www.hmpdacc.org/
4. Field, D. et al. *Nature Biotechnol* **26**, 541-547 (2008).
5. http://gensc.org
6. Liolios, K. et al. *Nucleic Acids Res* 36: D475-479 (2008).

7. For JGI members, work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396. The NMRC work was supported by grant TMTI0068_07_NM_T from the Joint Science and Technology Office for Chemical and Biological Defense (JSTO-CBD), Defense Threat Reduction Agency Initiative to TDR.