Opinion

Back to Bermuda: how is science best served?

Deanna M Church* and LaDeana W Hillier†

Addresses: *NCBI/NLM/NIH 8600 Rockville Pike, Bethesda, MD 20894, USA. †Department of Genome Sciences, University of Washington, 1705 NE Pacific Street, Seattle, WA 98195, USA.

Correspondence: Deanna Church. Email: church@ncbi.nlm.nih.gov

Published: 24 April 2009

Genome Biology 2009, **10:**105 (doi:10.1186/gb-2009-10-4-105)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/4/105

Abstract

The independent announcements of two bovine genome assemblies from the same data suggest it is time to revisit the spirit of the Bermuda and Fort Lauderdale agreements and determine the policies for data release and distribution that will best serve both the producers of the data and the users.

It is not possible to overstate the impact that genome sequencing and assembly has had on biomedical research. While the release of a new genome assembly once spawned worldwide press releases and announcements (in some cases multiple times) there is now a general expectation that if you are to do serious work on a model organism, a genome assembly is a necessary part of the research plan. These genome assemblies serve as the backbone for whole-genome studies, comparative genomics and for research labs performing locus-specific work. A critically important aspect of the success of the Human Genome Project (HGP) was the decision to immediately release pre-publication primary sequence data [1]. This policy flew in the face of tradition, especially in the community of those researching aspects of the human genome, which stated that genome sequence need only be made available upon publication. Although there was some concern that this would jeopardize the genome center's ability to analyze and publish the data they had produced, most involved felt that the benefit of early release outweighed the risks of an outside group publishing a genome assembly and analysis before the data producers. Guidelines for both the release and use of these data were published in what are commonly referred to as the Bermuda principles and the Fort Lauderdale agreement [2]. While the Bermuda principles have been incredibly valuable to the research community, they were established more than 10 years ago, and it is time to revisit them as sequencing technologies, standards and expectations are evolving at a rapid pace.

The necessity to revisit these guidelines is underscored by the simultaneous publication of two different assemblies of the cow genome: Btau 4.0 as described by the Bovine Genome Sequencing and Analysis Consortium (BGSAC) [3], and UMD 2.0 as described by Zimin *et al.* [4]. Both these genome assemblies are based on sequence traces generated by the Baylor College of Medicine as a part of the BGSAC. While the Zimin *et al.* publication does not violate the Fort Lauderdale agreement as both genomes are being published simultaneously, the availability of two genome assemblies produced from the same dataset raises a series of questions that will need to be addressed by funding agencies, sequence producers and the user community. How many assemblies are necessary and useful? Who has the right to perform the genome assembly? How should the community select reference assemblies? Are genome centers responsible for assembly updates forever?

Many users may be surprised that the same dataset would produce two different assemblies. However, the process of genome assembly is akin to putting together a 3 billion piece jigsaw puzzle. Of course, in the genome case many of the pieces look almost identical and there may be multiple correct solutions, depending on the data source. In addition to polymorphisms and alternative haplotypes, other complications include the presence of segmental duplications, defined as regions larger than 1 kb that have greater than 90% sequence identity with another region of the genome [5], and large-scale structural variation, meaning that two

chromosomes can differ by millions of base pairs or have regional ordering differences [6]. Even the two most complete and best studied mammalian genomes - human and mouse - which were produced by clone-based rather than whole-genome strategies, contain regions that remain unassembled or that contain errors [7].

Genome centers put a great deal of effort into producing high-quality sequence data and assemblies for the research community and they deserve to have the chance to assemble and analyze the data they produce. Although the effort involved in producing a genome assembly has not decreased, it is becoming increasingly difficult to get such work published. There is a danger that the effort required to perform the analysis required for publication in a top-tier journal can significantly delay publication of the genome. Whereas the assembly is typically available before publication, the inability of an outside group to publish a genome-wide analysis of an assembly before its publication can hinder the advancement of science. In other cases, there may be a substantial delay between the production of sequence reads and the production of the genome assembly. It is quite clear that the research community is not well served in these cases. It would be useful for the stakeholders to establish timelines by which such assembly and publication milestones should be reached.

A number of assembly programs are currently available but none produces a base-perfect assembly with data from current technologies. The shift from clone-based sequence to whole-genome sequencing and assembly (WGSA) means that the most highly duplicated, lineage-specific regions of the genome are poorly represented in the final assembly [8], but the way these regions are handled will vary with the assembly package. Because of complications like those described above, as well as the incomplete and non-uniform representation of the sequence in whole-genome sequencing datasets, even with a single assembly tool typically there are multiple possible solutions to any given assembly that are each completely consistent with the underlying data. Several projects have taken advantage of the fact that multiple assemblers exist and have produced multiple genome assemblies as a part of the project. For example, during the WGSA phase of the mouse genome projects, three rounds of assemblies were performed using two different genome assemblers (Arachne [9] and Phusion [10]). Both these assemblies were made available during the early stages of the project, but one was ultimately chosen for analysis and publication. A similar approach was taken for both the chimpanzee genome project [11] and the rhesus macaque genome project [12]. The availability of multiple algorithms and assemblies during the course of these projects improved the final product immensely. In all these projects the final assembly was made better because the different groups performing the assembly worked with the genome center responsible for the sequence data.

Everyone benefits if multiple assemblies are produced and compared. Statistics such as chromosome length and scaffold N50 (a measure of continuity that is defined as the scaffold length for which 50% of the bases in an assembly reside), although poor measures of base-level quality or global assembly correctness, are often taken into account when assessing assemblies. More importantly, comparison of the genome sequence to independently derived sequences, such as transcript collections or regions already finished using clone-based sequencing, has also proved an effective way to assess the quality of an assembly. Recently, additional approaches that look for inconsistencies in the assembled data have been described [13].

But despite the ability to perform many levels of analysis, there are typically no set metrics for determining which assembly should be deemed the reference. As different genomes have different biological characteristics and different levels of funding, it is difficult to establish a one-size-fits-all policy. However, at the beginning of each project it would be useful for all stakeholders to specify whether the analysis of multiple assemblies is desired and to define how any assemblies generated for the project will be measured. The development of a third-party group, perhaps consisting of representatives of the major annotator and browser groups, could assist the centers in the quality assessment stage of the assessment. Making the data from such assessments widely available, perhaps through the browsers, would help the user community understand both the positive aspects as well as the limitations of a given assembly. While it is generally advantageous to release a single assembly for a given dataset, there may be instances where it is not possible to determine the one best assembly, and in those cases it is better to release both.

There is an additional issue of assembly updates and improvements. Users performing genome-wide analysis want a single, stable coordinate system, whereas users interested in a specific gene or region want the best possible representation of that region. However, not all genome assemblies are updated after the initial publication. In many cases the centers no longer have funding to work on the projects, but the community continues to rely on the data and in many cases adds new data that could be used to improve the assembly. The resources generated by these large projects are too valuable to be allowed to lie fallow and we must explore mechanisms that do not burden the genome centers but enable the genome assembly to improve as our understanding of the data and genome increase. These may include continued funding to the center for the project or the transfer of the assembly to a third party for management and updates. This would be useful for the community as well as for the centers initially involved [7].

The notion of having multiple assemblies raises additional questions and underscores the need to develop better tools

for tracking, comparing and displaying genome-assembly data. As sequencing costs drop, additional datasets and assemblies will inevitably be produced. This is already the case for humans, for whom three different genome assemblies (the HGP public reference, Celera's, and Venter's) are already available. The overhead of analyzing, annotating and displaying genome sequences is considerable but manageable. However, the problems of data display, establishing stable coordinates for exchange and assembly tracking are considerable.

The first problem is assembly management. Although most assemblies are deposited in the International Nucleotide Sequence Database Consortium (INSDC) databases, commonly referred to as GenBank/EMBL/DDBJ, this is not sufficient for tracking the actual assembly, only the individual sequences associated with it. Currently, most assemblies are tracked by name and date, with no formal detailed notation of individual sequence changes. Tools for formally managing and tracking genome assemblies are currently in development, but they will only be the first step to the suite of tools that need to be developed for managing assemblies. There have been three updates to the human genome since the publication describing the 'finished' genome [14] and simply specifying that a feature is on human chromosome 1 at 10,000 base pairs is not sufficient to uniquely identify that base.

In addition to improved tools for tracking and managing assembly data, additional tools for comparing and displaying multiple assemblies need to be developed. Currently, Ensembl and the University of California Santa Cruz genome browser can only annotate and display a single current assembly within a given view, although archival versions of the reference assemblies are available. The National Center for Biotechnology Information has long supported the ability to annotate and display multiple assemblies for a given organism, but the book-keeping and user interface need improvement. Tools based on aligning assemblies and displaying comparative annotation are necessary to help most users navigate these data. In addition, tools for rapidly identifying assembly differences will be critical for honing in on regions that should be judged skeptically and may need manual intervention for improvement.

The sequencing of the human genome did not mark the end of sequencing, but merely the beginning. Sequence data are now easier to produce, but decisions about timelines for data release, publication, and ownership and standards for assembly comparison and quality assessment, as well as the tools for managing and displaying these data, need considerable attention in order to best serve the entire community.

References

1. **genome.gov | Policy on Release of Human Genomic Sequence Data (2000)** [http://www.genome.gov/page.cfm?pageID=10000910]
2. **genome.gov | February 2003 Data Release Policies** [http://www.genome.gov/10506537]
3. The Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC: **The genome sequence of taurine cattle: a window to ruminant biology and evolution.** *Science*, **324:**522-528.
4. Zimin AV, Delcher AL, Florea L, Kelley DA, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, Marcais G, Roberts M, Subramanian P, Yorke JA, Salzberg SL: **A whole-genome assembly of the domestic cow, *Bos taurus*.** *Genome Biol* 2009, **10:**r42.
5. Bailey JA, Eichler EE: **Primate segmental duplications: crucibles of evolution, diversity and disease.** *Nat Rev Genet* 2006, **7**:552-564.
6. Sharp AJ, Cheng Z, Eichler EE: **Structural variation of the human genome.** *Annu Rev Genomics Hum Genet* 2006, **7**:407-442.
7. **Genome Reference Consortium** [http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/]
8. She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL, Eichler EE: **Shotgun sequence assembly and recent segmental duplications within the human genome.** *Nature* 2004, **431:**927-930.
9. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES: **ARACHNE: a whole-genome shotgun assembler.** *Genome Res* 2002, **12:**177-189.
10. Mullikin JC, Ning Z: **The phusion assembler.** *Genome Res* 2003, **13**: 81-89.
11. The Chimpanzee Genome Sequencing Consortium: **Initial sequence of the chimpanzee genome and comparison with the human genome.** *Nature* 2005, **437:**69-87.
12. Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, Batzer MA, Bustamante CD, Eichler EE, Hahn MW, Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, Ding Y, *et al.*: **Evolutionary and biomedical insights from the rhesus macaque genome**. *Science* 2007, **316:**222-234.
13. Phillippy A, Schatz M, Pop M: **Genome assembly forensics: finding the elusive mis-assembly.** *Genome Biol* 2008, **9:**R55.
14. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431:**931-945.

Bovine genome coverage in BioMed Central:

- Burt DW: **The cattle genome reveals its secrets.** *J Biol* 2009, **8:**36.
- Capuco AV, Akers RM: **The origin and evolution of lactation.** *J Biol* 2009, **8:**37.
- Church DM, Hillier LW: **Back to Bermuda: how is science best served?** *Genome Biol* 2009, **10:**105.