U.S. DEPARTMENT OF ENERGY JOINT GENOME INSTITUTE

DOE JGI CONTACT: David Gilbert / 925.296.5643 / dgilbert@lbl.gov

Establishing Standard Definitions for Genome Sequences

WALNUT CREEK, CA—In 1996, researchers from major genome sequencing centers around the world convened on the island of Bermuda and defined a finished genome as a gapless sequence with a nucleotide error rate of one or less in 10,000 bases. This effectively set the quality target for the human genome effort and was quickly applied to other genome projects. If a genome sequence didn't meet this stringent criterion, it was simply considered a "draft."

More than a decade later, researchers are finding that with the advent of the latest sequencing technologies the terms "draft" and "finished" are no longer sufficient to describe the varying levels of genome sequence quality being produced. The quality issue is of particular concern for any researcher who wants to use the sequence, in order to know its integrity and reliability. This is of even greater concern for reference genome sequences, such as those genome projects conducted in support of the U.S. Department of Energy (DOE) missions of bioenergy and environmental clean-up, because they provide the foundational knowledge of the gene content and how these organisms interact with the environment.

As the proverbial "fire hose of data" becomes a Niagara torrent, with conservative estimates of 12,000 draft genomes hitting the public databases by 2012, researchers may be surprised to find that these datasets describe genomes that are not complete. Recognizing the problem, a group of researchers from several sequencing centers, including the DOE Joint Genome Institute (JGI), the Sanger Institute and the Human Microbiome Project (HMP) Jumpstart Consortium sequencing institutes, has proposed a new set of standards that expand upon the so-called "Bermuda standard." In the October 9 issue of the journal *Science*, they propose four additional categories between "draft" and "finished" status that reflect varying levels of completeness.

“In the past we’ve been limited to two options, requiring us and the other centers to come up with internal definitions,” said DOE JGI metagenomics researcher Patrick Chain at Los Alamos National Laboratory (LANL), first author of the *Science* paper. “But these are not clear and they’re not propagated to the databases to which we submit sequences. So when users try to download genomes they get data of unknown quality with no information, or a complete genome that they assume has been checked for missing-data errors.”

Chain said that when he and the other organizers of the Sequencing, Finishing, Analysis in the Future meeting hosted by LANL first gathered in 2005, they were concerned by the varying quality of the new genomes being submitted to public archives . As the meeting organizers all represented major sequencing centers (and smaller groups as well), the genome projects standards group was initiated at LANL, stimulated by these concerns.

The six categories defined by the group include:

- “Standard draft,” which is the minimum amount of information needed for submission to a public database;
- “High quality draft,” which is typically generated by large sequencing centers such as DOE JGI, and which has little or no manual review;
- “Improved high quality draft,” which consists of data reviewed by either people or machines to some extent so most of the genetic data is assembled correctly, but some errors may still be present;
- “Annotation-directed improvement,” which is a sequenced segment that presents all the information in various gene regions as accurately as possible;
- “Noncontiguous finished,” which includes sequences that have been reviewed by both people and machines and would be considered complete except for “recalcitrant regions” that are proving problematic;
- “Finished,” which defines complete sequences that have minimal errors, if any.

DOE JGI’s Chris Detter, one of the paper’s senior authors, and head of the LANL Genome Science group, said that the definitions provided in the *Science* paper are fairly flexible because

the group wanted the proposed standards to apply regardless of the genome project or sequencing technologies employed.

“My hope is all the major genome centers and advanced genomics groups use the gradations that fit their needs,” he said. “Some centers may want all six, while some may only want three, but as long as they keep them intact we are in good shape. Then, my hope is that the smaller genomics groups adopt the classes as written to help the rest of the scientific community know what they are generating and submitting.”

Chain added that the process of coming up with the proposed standards was not exactly an easy task since all major centers “have different pipelines, different sequencing techniques, different internal standards”. They also recognized that the attempt to develop a “one size fits all” set of standards is still a work in progress. The definitions provided in the *Science* paper are fairly flexible, designed to apply regardless of the genome project or sequencing technologies employed and to meet each group’s needs.

“We do expect that a number of people will comment on these standards, and possibly expand on the categories,” he said, “but we feel we’ve covered all the bases with these six categories.”

Chain said the group plans to team with the Genomic Standards Consortium, a grassroots movement begun by scientists who were concerned about the need for data collection standards in genome projects. The group has also talked to public archives such as GenBank to append these proposed standards to GenBank entries so that researchers can tell if the sequences will be useful to them. “Standards are a major issue to be tackled in genomics right now,” Chain said. “These proposals are guideposts meant to inform users and generators.”

Other DOE JGI authors on the study include David Bruce, Phil Hugenholtz, Nikos Kyrpides, Alla Lapidus, Sam Pitluck and Jeremy Schmutz. Other collaborating institutions are the Sanger Institute and the HMP Jumpstart Consortium sequencing centers (Washington University School of Medicine, the Broad Institute, the J. Craig Venter Institute, and Baylor College of Medicine), as well as Michigan State University, the Ontario Institute for Cancer Research, National Center

for Biotechnology Information, Seattle Children's Hospital and Research Institute, Emory GRA and the Naval Medical Research Center.

The U.S. Department of Energy Joint Genome Institute, supported by DOE's Office of Science, is committed to advancing genomics in support of DOE missions related to clean energy generation and environmental characterization and cleanup. DOE JGI, headquartered in Walnut Creek, Calif., provides integrated high-throughput sequencing and computational analysis that enable systems-based scientific approaches to these challenges. Follow DOE JGI on Twitter.

###