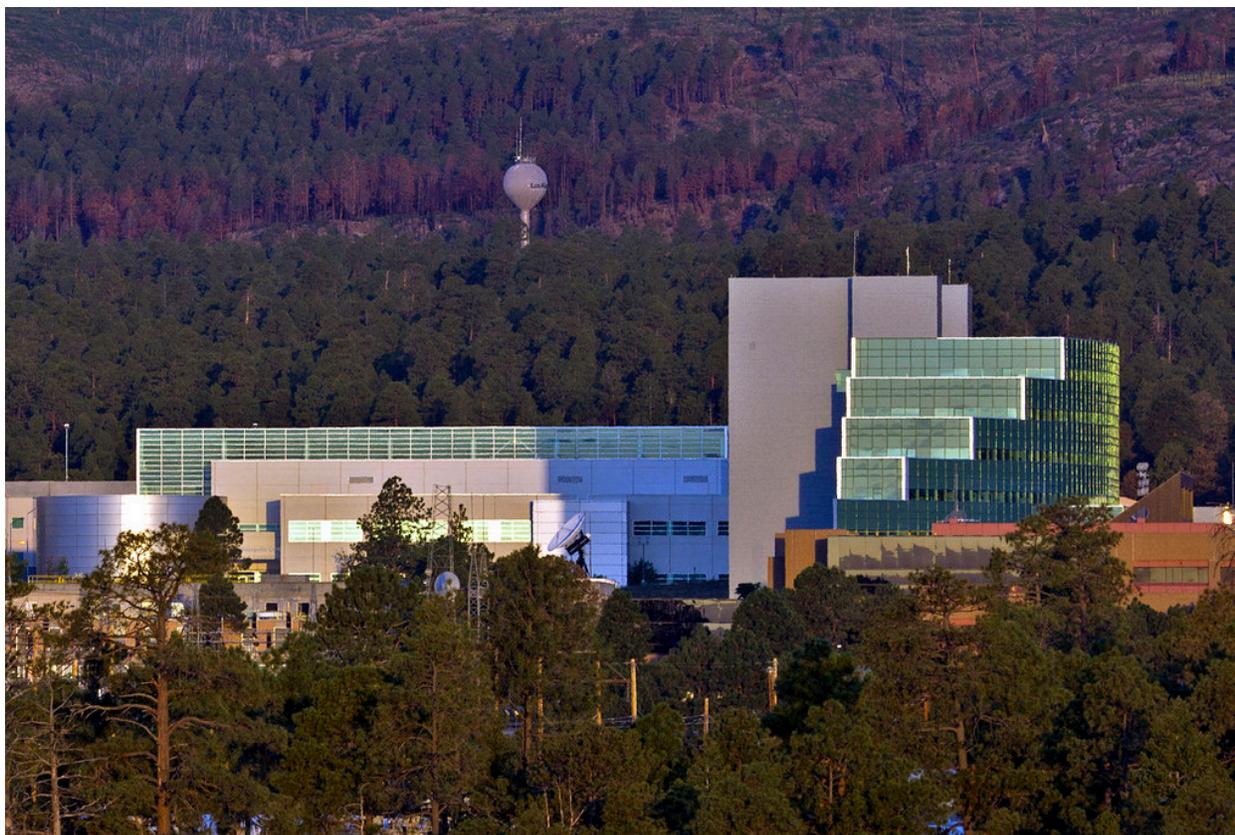


Standards for a new genomic era

October 21, 2009



LANL among organizations proposing new genome sequence strategies

Los Alamos, New Mexico, OCTOBER 21, 2009—A team of geneticists at Los Alamos National Laboratory, together with a consortium of international researchers, has recently proposed a set of standards designed to elucidate the quality of publicly available genetic sequencing information. The new standards could eventually allow genetic researchers to develop vaccines more efficiently or help public health or security personnel more quickly respond to potential public-health emergencies.

In a recent issue of *Science*, Los Alamos geneticist Patrick Chain and colleagues presented six labels for genome sequence data that are, or will become, available in public databases rather than the two labels used today. The six labels would roughly characterize the completeness and accuracy—and consequently, the potential reliability—of genetic sequencing data. This is of great importance since researchers use such

data on a daily basis for cross-referencing unknown genetic material with the genetic material of known organisms.

Every living organism with DNA has chromosomes containing the four molecular building blocks, or base pairs, represented by letters A, T, G, and C. One chromosome can contain millions of base pairs arranged like rungs on a ladder of DNA. The base pairs are arranged in sets of specific sequences that make up genes. These gene sequences can contain genetic instructions that help or harm an organism—for example by encoding enzymes that digest certain foods, or inducing cellular aberrations that give rise to certain diseases.

Genome researchers have catalogued genetic data from thousands of organisms and placed them in publicly available libraries. Researchers can use these libraries to crosscheck genetic data, for example when attempting to isolate an unknown public health threat, or to determine where a potentially helpful or harmful gene may be located on an organism's chromosome. For scientific fields such as biofuels research or environmental remediation, genetic data could help researchers determine whether microorganisms can efficiently break down plant matter to aid in ethanol production, or digest environmental contaminants like hydrocarbons.

However, because of the complexity of genetic data, genetic information in public libraries can range from very rough to very refined. In the past, genetic data has been classified either as “draft” or “finished,” leaving a wide range of uncertainty about the potential accuracy of genetic data.

“In the past few years we’ve seen major advances in genetic sequencing technology, so we’ve seen an explosion in the amount of publicly available data,” said Chain, who is lead author of the Science paper. “The amount of base-pair sequencing data generated each day is in the billions—orders of magnitude larger than what was generated a few years ago. Different sequencing technologies have different levels of accuracy. High degrees of uncertainty in a sequence can potentially lead a researcher down a wrong path that they could follow for a year or more. We now have a need for standards that will provide researchers with an unambiguous estimation of the quality of genetic sequence data.”

Working with researchers from genome sequencing centers big and small—including the U.S. Department of Energy's Joint Genome Institute, the Sanger Institute, the Human Microbiome Project Jumpstart Consortium sequencing centers, Michigan State University, and the Ontario Institute for Cancer Research, among others—Chain and colleagues have proposed that sequence data be placed into one of six categories that augment the existing two categories. The six standards range from “standard draft sequence,” representing minimum requirements for public submission, to a “finished sequence,” the highest standard, which can be verified to contain only one sequencing error per 100,000 base pairs.

“My hope is all the major genome centers and advanced genomics groups use the gradations that fit their needs,” said Chris Detter, LANL Genome Science Group Leader and Joint Genome Institute-LANL Center director. “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.”

Other DOE JGI authors on the Science paper 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.

Los Alamos National Laboratory

www.lanl.gov

(505) 667-7000

Los Alamos, NM

Operated by Los Alamos National Security, LLC for the Department of Energy's NNSA

