Defining genome project standards in a new era of sequencing

Patrick Chain

SFAF Meeting
Santa Fe, NM
May 2009
Welcome to our 4\textsuperscript{th} annual meeting!

- 1\textsuperscript{st} meeting (2006): Predictions regarding new technologies
  - Isolates, metagenomics…

- A ‘gap’ exists between generators and users
  - Sequences and Annotations
  - Similar gap between different sequence generators

- Standards should be implemented such that one can understand the type of information one can derive from a genome sequence
  - Genomic Standards Consortium – MIGS
  - Input and discussions at the first 3 meetings
Joint Announcement on Genome Sequencing Standards

- DOE Joint Genome Institute (LANL, LBNL, LLNL), Sanger Institute, Human Microbiome Project (WashU Genome Center, The Broad Institute, J. Craig Venter Institute, Baylor College of Medicine Human GSC, NHGRI, NIAID, U. of Maryland Institute for Genome Sciences), Emory U., HudsonAlpha GSC, Michigan State U., Ontario Institute for Cancer Research, NCBI, Naval Medical Research Center
  - Darren Grafham, Bob Fulton, Mike Fitzgerald, Donna Muzny, Jessica Hostetler, Chris Detter, many others...
1995: First complete genome of a free-living organism

1997: Second International Strategy Meeting on Human Genome Sequencing in Bermuda
  - [http://www.genome.gov/10001812](http://www.genome.gov/10001812)

Until very recently: genomes fell into 1 of 2 categories - draft or finished
  - Finished in bacteria/archaea is defined as base pair perfect sequence in a single contig
  - Finished in eukaryotes typically conforms to the bermuda standards, though most genomes are only regionally finished
  - This 2-category model requires significant re-evaluation…
2 categories: A growing problem

Genome Sequencing Projects on GOLD ©
Jan 2009, 4370 projects

- Incomplete
- Complete
Defining Genome Standards

A conservative projection: 12000 by 2012
Advances in sequencing technology

• With the advent of 2nd generation sequencing technologies, data generation has become easier and faster – large institutes and universities alike

• Will get worse:
  – PacBio; Helios; stay tuned...

Growth of GenBank (1982 - 2008)

Output per month/instrument

- capillary
  - 454
  - illumina
  - solid
New Technology for de novo Sequencing

• Replacing traditional whole genome Sanger sequencing
  – But still: gaps, repeats, misassemblies, homopolymers, SNPs, and other issues…

• Requires evaluation of methodology
  – Quality of: library, raw data, assemblies, combinations of technologies (and their applications)…

• Still two categories?
  – Finishing efforts need to be applied effectively and some definitions of gradations between draft and finished are required
2 categories: A growing problem

An agreement on gradation of finishing

- Finished
- Non-contiguous finished
- Annotation grade
- Improved high quality draft
- High quality draft
- Standard draft

Also, regionally improved…
An agreement on gradation of finishing

- Finished
- Non-contiguous finished
- Annotation grade
- Improved high quality draft
- High quality draft
- Standard draft

*may not always be possible to remove contaminating sequence data from this incomplete dataset*
An agreement on gradation of finishing

- Finished
- Non-contiguous finished
- Annotation grade
- Improved high quality draft
- High quality draft
- Standard draft

still a draft assembly (90% coverage) with little or no manual review of the product, thus sequence errors and misassemblies are possible, with no implied order and orientation to contigs
An agreement on gradation of finishing

Finished
Non-contiguous finished
Annotation grade
Improved high quality draft
High quality draft
Standard draft

Automated or manual improvement – with all but undetectable misassemblies addressed...low quality regions and potential bp errors may also be present, but the sequence is of high quality
Efforts have been made to resolve all errors in gene regions (exceptions to this gene-specific finishing standard should be noted with comments) ... repeat regions are not necessarily resolved, so errors in those regions are much more likely.
An agreement on gradation of finishing

Finished
Non-contiguous finished
Annotation grade
Improved high quality draft
High quality draft
Standard draft

All gaps and sequence uncertainties have been attempted to be resolved, and only those recalcitrant to resolution remain, but are specifically noted as to the nature of the uncertainty.
An agreement on gradation of finishing

- Standard draft
- High quality draft
- Improved high quality draft
- Annotation grade
- Non-contiguous finished
- Finished

Gold Standard (1 error per 100,000 bp, no misassemblies).
Any remaining small exceptions to highly accurate sequence are commented in the submission.
One system for all

- Agreed upon by all members of our consortium
- Will vastly improve user understanding
- Planned article in journal and publicise on institute webpages
- Discussions with databases underway
- Technology-agnostic so can adapt as 3rd (next) generation of sequencers arrives
- Project/centres can add more detail as required such as sequence coverage and platform type
- Integrates well with other domains of the GSC
  - SIGS (www.standardsingenomics.org) and MIGS