Towards a Library of Standard Operating Procedures (SOPs) for (meta)genomic annotation

Samuel V Angiuoli[1], Guy Cochrane[2], Dawn Field[3], George Garrity[4], Aaron Gussman[1], Chinnappa D Kodira[5], William Klimke[6], Nikos Kyrpides[7], Ramana Madupu[8], Victor Markowitz[7], Tatiana Tatusova[6], Nick Thomson[9], Owen White[1]

[1] University of Maryland

[2] EMBL-EBI

[3] Oxford

[4] University of Michigan

[5] Broad

[6] NCBI

[7] JGI

[8] JCVI

[9] Sanger

Abstract

Genome annotations describe the features of genomes and accompany sequences in genome databases. The methodologies used to generate genome annotation are diverse and typically vary amongst groups. Descriptions of the annotation procedure are helpful in interpreting genome annotation data. Standard Operating Procedures (SOPs) for genome annotation describe the processes that generate genome annotations. Some groups are currently documenting procedures but standards are lacking for structure and content of annotation SOPs. In addition, there is no central repository to store and disseminate procedures and protocols for genome annotation. We highlight the importance of SOPs for genome annotation and endorse a central online repository of SOPs.

Introduction

Genome annotation involves processes during which genome sequences are annotated with biological features, such as genes and proteins, and descriptors, such as gene names and protein functions. We define genome annotation broadly to encompass electronic information about various types of genomic data, including whole genome sequence data and metagenomic sequence data. Genome sequencing centers regularly produce genome annotation data in addition to producing raw sequence data in the form of sequencing reads and assemblies. In addition, many consumers of sequence data, such as online databases and resources, generate additional genome annotations that supplement those produced by the sequencing center (see for example (Sterk, Kersey et al. 2006; Flicek, Aken et al. 2008)). While many such resources provide direct public access to their supplementary annotations, the public nucleotide databases of the INSDC are also able to present some such data (Benson, Karsch-Mizrachi et al. 2008; Cochrane, Akhtar et al. 2008; Sugawara, Ogasawara et al. 2008) .

Genome biologists and bioinformaticists employ numerous computational tools to generate annotation about genomes and genes. Some annotation pipelines are based on sequence homology, using tools such as BLAST(Altschul, Gish et al. 1990), and are sensitive to parameters or applied cutoffs that can affect outcomes. Often the results of multiple tools are combined as evidence for a single annotation. Additionally, annotation processes may include curatorial steps where domain experts perform quality assessments and make decisions that affect the process flow and final annotation. Yet, in the public sequence databases and online resources, full descriptions of the procedures used to combine or derive evidence for an annotation are not regularly available. In some cases, a description of the annotation procedure may appear in an associated publication or project web site, but these descriptions may not be sufficient to reproduce the pipeline or determine the exact procedures that produced a specific annotation.

Standard operating procedures (SOPs) are human-readable documents that describe steps of a process and are widely adopted in many disciplines where it is important that a process is repeatable or auditable. The Genome Standard Consortium (GSC) is an organization promoting standards that increase the richness and usability of genomic datasets(Brooksbank and Quackenbush 2006; Field, Morrison et al. 2006). As representative of the GSC, we promote documentation of SOPs for genome annotation as a way to increase transparency and quality of the annotation process. SOPs are complementary to the minimal data standard efforts where

SOPs describe processes which generate data sets rather than dictate elements of a minimal data set.

What is a genome annotation SOP?

A genome annotation SOP describes processes used to generate annotations about a genomic sequence. The SOP should list the input and outputs of the process, reference any external tools used, such as software packages, and describe the primary steps of the process in detail. An annotation SOP will often include a combination of computational (automated) or curatorial (manual) steps of a data generation or data analysis procedure. The SOP should be described in sufficient detail such that a domain expert could replicate the annotation process using the appropriate tools. It is particularly important that an SOP also describe any evaluation points or quality assurance steps of a process in detail because often these steps are critical for understanding or replicating a process. For example, a quality assurance step of an SOP can detail conditions when the results of a particular computational analysis are trusted or discarded.

In this paper, we concern ourselves with large-scale genome and metagenome sequencing projects. However, we recognize that a great deal of annotation data exists, and will continue to be generated, as part of small-scale studies of fragmented nucleotide sequences from isolated organisms and environmental sampling. While we intend that SOP reporting conventions that might be established as part of this initiative will inform future developments in small-scale annotation reporting, such annotations suffer less from poor quality than their large-scale counterparts; small-scale data are, by nature, submitted as part of small studies, in which the literature references focus with great intensity upon the annotation presented and the approach through which it was generated (unlike large-scale annotation, where specific annotation objects are rarely mentioned in associated literature) and small-scale data typically reach the public domain through submission to INSDC databases using web-based tools and direct communication with database curators to optimize annotation, leading in particular to extensive and sophisticated use of evidence code structures.

We make note that for computational processes, a mere list of software and parameters is usually not sufficient to describe a process. The SOPs should include a description of how the outputs of software packages are interpreted, filtered, or combined with other outputs. We recognize that annotation pipelines may include numerous software packages that have a complex set of embedded rules or that function as an opaque “black box”. Although SOPs are intended to make the steps of a pipeline more transparent, an annotation SOP need not enumerate all the conditions and rules that are embedded within software. Rather, the SOP should describe how to use a software system so that another user of the system could be expected to generate a compatible result.

Several protocols for varying types of genome annotation are available online at web sites for genome annotation centers. Table 1 provides URLs to some annotation SOPs currently available on the Internet. Some of the SOPs in this list were produced through coordinated efforts that have recognized and promoted the publication of annotation SOPs(Greene, Collins et al. 2007). A review of these SOPs shows a diversity of scopes, content, and syntax.

Why are SOPs important for genome annotation?

SOPs help evaluate genome annotation data. It is currently difficult to trace the processes that are used to produce genome annotations. For example, users of genome annotation cannot always readily distinguish between annotations that are produced by purely computational methods and those reviewed by expert curators(Kyrpides and Ouzounis 1999). This problem has been recognized by groups such as the Gene Ontology consortium (Ashburner, Ball et al. 2000) and the INSDC, who provide evidence codes for referencing annotation methods. Gene Ontology consortium examples of evidence codes include IEA, "Inferred from Electronic Annotation" and ISS, "Inferred from Sequence Similarity", both of which can be combined with references to supporting evidence, such as a literature citation or an accession in a sequence database(GO). INSDC examples include '/inference="ab initioprediction:Genscan:2.0"', '/inference="similar to DNA sequence:INSD:AY411252.1"' and '/experiment="heterologous expression system of Xenopus laevis oocytes"'. Importantly, evidence codes do not attempt to describe the entire process or set of decisions that led to a particular annotation, rather they attempt to present specific information that relates the annotation in question to objects (literature, database records, tools) that specifically impacted on their generation. For these reasons, we see SOPs as a complementary effort to using evidence codes for annotations. SOPs describe the process that resulted in the assignment of a particular evidence code and supporting evidence.

SOPs help users of genome data understand inconsistencies between annotations produced by different methodologies. Numerous genome annotation pipelines have lead to heterogeneity in genome annotation databases (Brenner 1999; Devos and Valencia 2001). Comparisons of annotation pipelines have recognized conflicting gene annotations from pipelines that utilize similar tools or follow similar principles (Kyrpides and Ouzounis 1999; Iliopoulos, Tsoka et al. 2003; Tetko, Brauner et al. 2005). In addition, genome annotations in public databases are fraught with errors(Brenner 1999). SOPs don't directly provide a way to resolve heterogeneity or errors in genome databases. But, by describing the process, SOPs can help users of genome data understand reasons for inconsistent or erroneous outcomes. In contrast, without SOPs, users are left with little explanation as to why particular annotations are present or absent from a data set.

SOPs facilitate the exchange of process descriptions amongst domain experts who are interested in improving annotation quality. Comparisons of annotation processes have recognized challenges in assessing annotation quality(Tetko, Brauner et al. 2005). By making the annotation process more transparent, SOPs aid in the evaluation of competing systems, which can help propel improvements to the state of the art across the community.

An online library of SOPs

We propose development of a centralized, online electronic repository as a library for storing genome-scale annotation SOPs. A central online repository will simplify access to SOPs and facilitate searching and comparisons of SOPs. One model for an online repository is an open access electronic journal, where SOPs are submitted as publications. Other models for electronic repositories include web sites, such as a wiki site, where users can directly upload or edit their SOPs. Any successful model should allow the submitters of SOPs to update and modify them over time, applying appropriate version tracking systems. An advantage to treating SOPs as journal publications is that the SOPs can then be cited in the scientific literature. The publishing model also provides for a review process where SOPs may be reviewed for syntax and structure prior to publication to ensure a level of quality.

We propose that unique identifiers with version numbers are assigned to SOPs. Unique identifiers simplify external linking to SOPs on the World Wide Web. Furthermore, unique identifiers provision for associating annotation outcomes and SOPs in genome databases. In one scenario, genome annotations are tagged with SOP identifier(s) signifying the processes that produced the annotation. By linking SOPs identifiers to annotation objects, such as genes and their names, users of genome databases will be able to better track the processes used to generate the annotations. One model to achieve this would be to encourage the submitters of genome annotations to the INSDC to publish their SOPs in the central repository prior to submission and then to provide links to these SOPs as part of the submission of genome annotation. A central repository for SOPs will be responsible for providing a mechanism to assign unique identifiers. We note that the publication model already provides a standard for creating stable links and unique identifiers for documents using Digital Object Identifiers (DOIs).

Formats for annotation SOPs

A common syntax for SOPs follows a semi-structured document with numbered heading and subheadings, such as 1.1 Title, 1.2 Overview, 2.1 Procedures. Other formats include unstructured and narrative text, or a highly structured document such as XML with a DTD. One advantage of structured documents and a DTD is that they can easily be parsed by computers. Any suggested format for SOPs should encourage submission of process details. Less structured text, such as is typically found in the methods section of a paper, often lack the detail required to trace and fully replicate an annotation process.

The annotation SOPs currently on the web, such as those in Table 1, are diverse in format and document structure. A central repository should promote a standard format(s) to aid creation and dissemination of SOPs. We recommend that a standard format accommodate SOPs written at varying levels of detail. SOPs should include basic administrative elements such as a title, author(s), institute(s) of origin, a revision version and date. An SOP should provide a brief text overview (or abstract) describing the SOP and a category listing the type of annotation process described.

An annotation SOP case study

As a case study, we provide a excerpt from a SOP (http://www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html) that generates draft annotation of complete prokaryotic genomes(Daraselia, Dernovoy et al. 2003). The process named the Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP) follows in a narrative format. "The PGAAP combines Hidden Markov Model (HMM) based gene prediction methods with a sequence similarity-based approach which combines comparison of the predicted gene products to the non-redundant protein database, Entrez Protein Clusters(Wheeler, Barrett et al. 2008) , the Conserved Domain Database(Marchler-Bauer, Anderson et al. 2005), and the Clusters of Orthologous Groups (COGs)(Tatusov, Fedorova et al. 2003). Submitters requesting the use of the annotation pipeline for their genomic sequences submit them to NCBI in FASTA format. Gene predictions are done using a combination of GeneMark(Borodovsky and McIninch 1993;

Lukashin and Borodovsky 1998) and Glimmer(Salzberg, Delcher et al. 1998). A short step resolving conflicts of start sites is done at this point. Ribosomal RNAs are predicted by sequence similarity searching using BLAST(Altschul, Gish et al. 1990) against an RNA sequence database and,or using Infernal and Rfam models(Griffiths-Jones, Moxon et al. 2005). Transfer RNAs are predicted using tRNAscan-SE(Lowe and Eddy 1997). In order to detect missing genes, a complete six-frame translation of the nucleotide sequence is done and predicted proteins (generated above) are masked. All predictions are then searched using BLAST against all proteins from complete microbial genomes. Annotation is based on comparison to protein clusters and on the BLAST results. Conserved Domain Database and Cluster of Orthologous Group information is then added to the annotation. Frameshift detection and cleanup occurs and then the final output is then sent back to the submitters who can then analyze the results in preparation for submission to GenBank."

This SOP provides a general description of an annotation pipeline and a motivating example of an annotation SOP. A comparison of the SOPs in Table 1 show varying levels of detail in describing protocols. Relevant software parameters or cutoffs and detailed descriptions of quality assurance steps are important elements of processes but are not described fully in all the available SOPs.

Conclusion

SOPs stand to improve understanding of genome annotations and clarify an often opaque process. SOPs also provide a good starting point for advocating and improving best practices across the genome annotation community. We seek SOPs of the detail required to allow for precise replication of annotation pipelines. But, we also recognize that writing SOPs that allow for reproducibility is neither easy nor always practical. Documentation of protocols is laborious and requires extensive domain expertise. We seek a document format that simplifies documenting annotation protocols.

Existing protocols published on genome annotation web sites show a diversity of content and format. We embrace a diversity of annotation protocols and recognize an opportunity to create a centralized repository for SOPs. We see an online repository of SOPs as an important resource for members of the genome annotation community. The electronic journal and publication model with a baseline review process is an intriguing model for an online annotation SOP repository.

Titles or scope	URL
NCBI Prokaryotic Genomes Automatic Annotation Pipeline	http://www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html http://www.ncbi.nlm.nih.gov/genomes/static/Annotation_pipeline_procedures.txt
Gene prediction, protein product assignment	http://img.jgi.doe.gov/pub/doc/img_er_ann.pdf
Gene structure prediction, gene naming, quality control	http://www.broad.mit.edu/annotation/genome/batrachochytrium_dendrobatidis/GeneFinding.html http://www.broad.mit.edu/annotation/genome/francisella_tularensis_group/GeneFinding.html
Gene Curation, Analysis and Curation of Short Gene Models, Homology Searches, Functional Automated Annotation, Functional Manual Curation, Start Site Curation, Frameshift Edit and Analysis, Overlap Analysis and Curation	http://cmr.jcvi.org/CMR/TigrAnnotationSops.shtml
Genomic Sequence Annotation Pipeline, Automated DNA-Level Curation, Manual DNA-Level Feature Curation, Protein Annotation Pipeline, Automated Protein Curation Pipeline, Orthologous Gene Predition	http://patric.vbi.vt.edu/about/standard_procedures.php
CDS Annotation, Ortholog Assignment and Curation, Annotation of Insertion Sequences, Pseudogene Annotation, RNA Gene Annotation, Polymorphism Annotation	http://www.ericbrc.org/portal/eric/aboutasap
Automated Annotation	http://www.biovirus.org/docs.asp#publications
Gene structure inferred from protein and transcript data	http://www.vectorbase.org/Help/Category:VectorBase_SOP
Gene Model and Functional Curation	http://cryptodb.org/static/SOP/

Table 1: SOPs related to genome annotation currently available on the web

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Benson, D. A., I. Karsch-Mizrachi, et al. (2008). "GenBank." Nucleic Acids Res **36**(Database issue): D25-30.

Borodovsky, M. and J. McIninch (1993). "Recognition of genes in DNA sequence with ambiguities." Biosystems **30**(1-3): 161-71.

Brenner, S. E. (1999). "Errors in genome annotation." Trends Genet **15**(4): 132-3.

Brooksbank, C. and J. Quackenbush (2006). "Data standards: a call to action." OMICS **10**(2): 94-9.

Cochrane, G., R. Akhtar, et al. (2008). "Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database." Nucleic Acids Res **36**(Database issue): D5-12.

Daraselia, N., D. Dernovoy, et al. (2003). "Reannotation of Shewanella oneidensis genome." OMICS **7**(2): 171-5.

Devos, D. and A. Valencia (2001). "Intrinsic errors in genome annotation." Trends Genet **17**(8): 429-31.

Field, D., N. Morrison, et al. (2006). "Meeting report: eGenomics: Cataloguing our Complete Genome Collection II." OMICS **10**(2): 100-4.

Flicek, P., B. L. Aken, et al. (2008). "Ensembl 2008." Nucleic Acids Res **36**(Database issue): D707-14.

GO. "Guide to GO evidence codes." from http://www.geneontology.org/GO.evidence.shtml.

Greene, J. M., F. Collins, et al. (2007). "National Institute of Allergy and Infectious Diseases bioinformatics resource centers: new assets for pathogen informatics." Infect Immun **75**(7): 3212-9.

Griffiths-Jones, S., S. Moxon, et al. (2005). "Rfam: annotating non-coding RNAs in complete genomes." Nucleic Acids Res **33**(Database issue): D121-4.

Iliopoulos, I., S. Tsoka, et al. (2003). "Evaluation of annotation strategies using an entire genome sequence." Bioinformatics **19**(6): 717-26.

Kyrpides, N. C. and C. A. Ouzounis (1999). "Whole-genome sequence annotation: 'Going wrong with confidence'." Mol Microbiol **32**(4): 886-7.

Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." Nucleic Acids Res **25**(5): 955-64.

Lukashin, A. V. and M. Borodovsky (1998). "GeneMark.hmm: new solutions for gene finding." Nucleic Acids Res **26**(4): 1107-15.

Marchler-Bauer, A., J. B. Anderson, et al. (2005). "CDD: a Conserved Domain Database for protein classification." Nucleic Acids Res **33**(Database issue): D192-6.

Salzberg, S. L., A. L. Delcher, et al. (1998). "Microbial gene identification using interpolated Markov models." Nucleic Acids Res **26**(2): 544-8.

Sterk, P., P. J. Kersey, et al. (2006). "Genome Reviews: standardizing content and representation of information about complete genomes." OMICS **10**(2): 114-8.

Sugawara, H., O. Ogasawara, et al. (2008). "DDBJ with new system and face." Nucleic Acids Res **36**(Database issue): D22-4.

Tatusov, R. L., N. D. Fedorova, et al. (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics **4**: 41.

Tetko, I. V., B. Brauner, et al. (2005). "MIPS bacterial genomes functional annotation benchmark dataset." Bioinformatics **21**(10): 2520-1.

Wheeler, D. L., T. Barrett, et al. (2008). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **36**(Database issue): D13-21.