OMICS A Journal of Integrative Biology
Volume 12, Number 2, 2008

DOI: 10.1089/omi.2008.0017

Toward an Online Repository of Standard Operating Procedures (SOPs) for (Meta)genomic Annotation

Samuel V. Angiuoli,[1,2] Aaron Gussman,[1] William Klimke,[3] Guy Cochrane,[4] Dawn Field,[5] George M. Garrity,[6] Chinnappa D. Kodira,[7] Nikos Kyrpides,[8] Ramana Madupu,[9] Victor Markowitz,[10] Tatiana Tatusova,[3] Nick Thomson,[11] and Owen White[1]

Abstract

The methodologies used to generate genome and metagenome annotations are diverse and vary between groups and laboratories. Descriptions of the annotation process are helpful in interpreting genome annotation data. Some groups have produced Standard Operating Procedures (SOPs) that describe the annotation process, but standards are lacking for structure and content of these descriptions. 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 an online repository of SOPs.

Introduction

Genome annotation involves processes during which genome sequences are marked up with descriptive notations, such as names and functions, about known or postulated biological features (Stein, 2001). Genome annotation could be defined even more broadly to encompass any electronic information about various types of genomic data, including whole genome sequence data and metagenomic sequence data. The general scientific community is presented with annotation that comes from a variety of sources. Genome sequencing centers regularly produce genome annotation with primary sequence information. In addition, online resources, such as the NIAID Bioinformatics Resource Centers (BRCs) (Greene, 2007), CMR (Peterson et al., 2001), IMG (Markowitz et al., 2008), and Ensembl (Flicek et al., 2008), generate and display additional genome annotations. The public nucleotide databases of the International Nucleotide Sequence Database Collaboration (INSDC) are also able to incorporate some of these annotations (Benson et al., 2008; Cochrane et al., 2008; Sugawara et al., 2008).

The genome annotations in public resources are useful, but unfortunately, the methods used to generate the data are not obvious to the biologist using this data. This is because there are a number of different processes that may be employed to generate annotation about genomes. Some annotation pipelines are based on sequence homology, using tools such as BLAST (Altschul 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.

[1]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland.
[2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland.
[3]National Center for Biotechnology Information, National Institutes of Health, Baltimore, Maryland.
[4]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom.
[5]Molecular Evolution & Bioinformatics Section, Mansfield Road, Oxford Centre for Ecology and Hydrology, Oxford, United Kingdom.
[6]Department of Microbiology and Molecular Genetics at Michigan State University, East Lansing, Michigan.
[7]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts.
[8]Department of Energy, Joint Genome Institute, Walnut Creek, California.
[9]J. Craig Venter Institute, Rockville, Maryland.
[10]Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, California.
[11]The Pathogen Sequencing Unit, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom.

What is a Genome Annotation SOP?

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 need for SOPs describing the genome annotation processes was highlighted at the fifth workshop of the Genomic Standards Consortium (Field et al., 2008a, 2008b). In its capacity as an organization to promote standards that increase the richness and usability of genomic datasets (Field et al., 2008a), the GSC advocates that SOPs for genome annotation become more routinely used, as a way to increase transparency and quality of the annotation process.

Table 1 provides links to some of the annotation SOPs currently available on the Internet. These SOPs are diverse in scope, content, and syntax and highlight the need for standardization. Some of the SOPs in this list were produced through coordinated efforts that have recognized and promoted the publication of annotation SOPs (Greene et al., 2007).

Genome annotation SOPs should accomplish a number of tasks. They should document specific processes used to generate annotations about a genomic sequence. Each SOP should list the input and outputs of the annotation process, reference any external tools or software that were used 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 annotation procedure should be described in sufficient detail such that a domain expert (bioinformatician) could replicate the anno-

Table 1. SOPs Related to Genome Annotation Currently Available on the Web

Titles or scopes	*Project or center*	*URL*
NCBI prokaryotic genomes automatic annotation pipeline	NCBI	http://www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html
Gene prediction, protein product assignment	JGI	http://img.jgi.doe.gov/pub/doc/img_er_ann.pdf
Gene structure prediction, gene naming, quality control	Broad Institute	http://www.broad.mit.edu/annotation/fgi/GeneFinding.html http://www.broad.mit.edu/seq/msc/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	JCVI	http://cmr.jcvi.org/CMR/TigrAnnotationsSops.shtml
Genomic sequence annotation pipeline, automated DNA-level curation, manual DNA-level feature curation, protein annotation pipeline, automated protein curation pipeline, orthologous gene prediction	PATRIC (Synd et al., 2007)	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	ERIC (Glasner, et al., 2008)	http://www.ericbrc.org/portal/eric/aboutasap
Automated annotation	VBRC	http://www.biovirus.org/docs.asp#publications
Gene structure inferred from protein and transcript data	Vectorbase (Lawson, et al., 2007)	http://www.vectorbase.org/Help/Category:VectorBase_SOP
Gene model and functional curation	Cryptodb (Heiges et al., 2006)	http://cryptodb.org/static/SOP/

tation process using the listed tools. The SOPs should include a description of how the outputs of software packages are interpreted, filtered, or combined with other outputs. We note that SOPs are not simply a list of software and parameters. It is particularly important that SOPs also describe evaluation points or quality assurance steps of a process in detail because often these are critical for understanding or replicating a process. For example, quality assurance steps of an SOP can describe when the results of particular computational analyses are trusted or discarded. We recognize that annotation pipelines may include numerous software packages that have a complex set of embedded rules or that function as a "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. The SOP should describe how to use a software system so that another user of the system could expect to generate a compatible result.

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. 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. While process descriptions from small scale studies can benefit from the added structure of SOPs, smaller scale studies may not lend themselves to SOPs as much as large-scale high-throughput sequencing projects.

Why are SOPS Important for Genome Annotation?

SOPs help users (biologists and bioinformaticians) evaluate the nature (quality and quantity) of genome annotation data. It is currently difficult to trace the processes that are used to produce genome annotations. For example, users of genome annotations cannot always readily distinguish between those 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 et al., 2000) and the INSDC, both of which 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. INSDC examples include /inference="ab initio prediction: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 references to objects (literature, database records, tools) that specifically impacted the annotation. 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 among annotations produced by different methodologies. Numerous genome annotation pipelines have lead to heterogeneity in genome annotation databases. Comparisons of annotation pipelines have shown conflicting gene annotations from pipelines that utilize similar tools or follow similar principles (Kyrpides and Ouzounis 1999; Iliopoulos, Tsoka et al. 2003; Tetko et al., 2005). In addition, genome annotations in public databases are fraught with errors particularly for functional annotations (Brenner, 1999; Devos and Valencia, 2001; Valencia, 2005). SOPs do not 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. Another benefit is that SOPs facilitate the exchange of process descriptions among domain experts who are interested in improving annotation quality. 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.

A Centralized Online Repository of SOPs

We propose development of a centralized, online repository as a library for storing genome annotation SOPs. Such a repository will simplify access to SOPs and facilitate searching and comparisons of SOPs. One model for an online repository is to create a new form of Open Access electronic journal, where SOPs are submitted as a type of formal publication (Garrity et al., 2008). Other models for such an electronic repository include a single Web site maintained by a single group or a Wiki site, for example maintained by the GSC, where users can directly upload or edit their SOPs. Any successful model adopted in the long term 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. Finally, it provides a way for downstream use of a particular SOP, or a modification of it, to cite it, and in this way spread a best-practice throughout the wider community.

Linking annotations to SOPs through unique identifiers

We propose SOPs be assigned unique persistent identifiers with version numbers. Unique identifiers provide a mechanism to link to SOPs on the World Wide Web. More importantly, unique identifiers also provide a means of associating annotation outcomes with SOPs in genome databases. In this scenario, genome annotations are tagged with SOP identifier(s) identifying the processes that produced the annotation allowing users to track the processes used to generate the annotations. One model to achieve this would encourage the submitters of genome annotations to the INSDC to publish their SOPs in the central repository prior to submission and provide links to their SOPs as part of the submission. We note that the publication industry already pro-

vides one standard for creating stable links and unique identifiers for documents using Digital Object Identifiers (DOIs) (Paskin, 2005).

Formats for Annotation SOPs

The annotation SOPs currently on the Web, such as those in Table 1, are diverse in format and document structure. Many follow a semistructured document format with numbered heading and subheadings, such as 1.1. Title, 1.2. Overview, 2.1. Procedures. A central repository should promote a standard format(s) for SOPs that defines required and recommended elements. SOPs should also 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 of the type of annotation process described. The format should encourage submission of process details and provide mechanisms for the documentation of software invocation parameters and cutoffs. One mechanism to promote detail is to utilize a highly structured format and define a Document Type Definition (DTD), although none of existing annotation SOPs in Table 1 utilize a DTD. Structured documents and DTDs ensure consistency and standardization of content and can be easily parsed by computers for searching and querying.

An Annotation SOP Case Study

As a case study, we provide an excerpt from an SOP (http://www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html) that generates a draft annotation of a complete prokaryotic genomes (Daraselia 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 nonredundant protein database, Entrez Protein Clusters (Wheeler et al., 2008), the Conserved Domain Database (Marchler-Bauer et al., 2005), and the Clusters of Orthologous Groups (COGs) (Tatusov 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 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 et al., 1990) against an RNA sequence database and/or using Infernal and Rfam models (Griffiths-Jones et al., 2005). Transfer RNAs are predicted using tRNAscan-SE (Lowe and Eddy, 1997). 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 representative example of an annotation SOP. Like many SOPs in Table 1, this SOP is described at a high level and does not fully describe software parameters, cutoffs, or quality assurance steps. These details are important elements of an SOP if the procedure is to be easily reproducible. We see this case study SOP as a good first step toward documenting an annotation process. A standard SOP format that defines requirements and an associated central repository with exemplar SOPs should help promote generation of SOPs that allow for better comparability and reproducibility.

Conclusion

SOPs improve end-users' understanding of genome annotations and clarify an often opaque process of genome annotation. SOPs also provide a good starting point for advocating and improving best practices across the genome annotation community. We seek SOPs with sufficient required detail 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 SOP format that simplifies documenting annotation protocols and standardizes content.

We embrace the 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 are intriguing models for an online annotation SOP repository.

Acknowledgment

This work has been supported by the United States National Institute of Allergy and Infectious Disease Contract NIH-NIAID-DMID-04-34, HHSN2662004000386.

Author Disclosure Statement

The authors declare that no competing financial interests exist.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol **215,** 403–410.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, D.M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25,** 25–29.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2008). GenBank. Nucleic Acids Res **36,** D25–D30.

Borodovsky, M., and Mcininch, J. (1993). Recognition of genes in DNA sequence with ambiguities. Biosystems **30,** 161–171.

Brenner, S.E. (1999). Errors in genome annotation. Trends Genet **15,** 132–133.

Cochrane, G., Akhtar, R., Aldebert, P., Althorpe, N., Baldwin, A., Bates, K., et al. (2008). Priorities for nucleotide trace, se-

quence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. Nucleic Acids Res **36,** D5–D12.

Daraselia, N., Dernovoy, D., Tian, Y., Borodovsky, N., Tatusov, R., and Tatusova, T. (2003). Reannotation of *Shewanella oneidensis genome*. OMICS **7,** 171–175.

Devos, D., and Valencia, A. (2001). Intrinsic errors in genome annotation. Trends Genet **17,** 429–431.

Field, D., Garrity, G.M., et al. (2008a). Towards a richer description of our complete collection of genomes and metagenomes: the minimal information about a genome sequence. Nat Biotechnol **26,** 541–547.

Field, D., Glöckner, F.O., et al. (2008b). Meeting report: The 4th genomic standards consortium (GSC) workshop. OMICS (this issue).

Flicek, P., Aken, B.L., Beal, K., Ballester, B., Caccano, M., Chen Y., et al. (2008). Ensembl 2008. Nucleic Acids Res **36,** D707–D714.

Glasner, J.D., Plunkett, G., III, Anderson, B.D., Baumler, D.J., Bichl, B.S., Burland, V., et al. (2008). Enteropathogen Resource Integration Center (ERIC): bioinformatics support for research on biodefense-relevant enterobacteria. Nucleic Acids Res **36,** D519–D523.

Greene, J.M., Collins, F., Lefkovitz, E.J., Roos, D., Scheuermann, R.H., Sobral, B., et al. (2007). National Institute of Allergy and Infectious Diseases bioinformatics resource centers: new assets for pathogen informatics. Infect Immun **75,** 3212–3219.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res **33,** D121–D124.

Heiges, M., Wang, H., Robinson, E., Aurrecoechea, C., Gao, X., Kaluskar, N., et al. (2006). CryptoDB: a *Cryptosporidium* bioinformatics resource update. Nucleic Acids Res **34,** D419–D422.

Iliopoulos, I., Tsoka, S., Andrade, M.A., Enright, A.J., Caroll, M., Poullet, P., et al. (2003). Evaluation of annotation strategies using an entire genome sequence. Bioinformatics **19,** 717–726.

Kyrpides, N.C., and Ouzounis, C.A. (1999). Whole-genome sequence annotation: "Going wrong with confidence." Mol Microbiol **32,** 886–887.

Lawson, D., Arensburger, P., Atkinson, P., Besansky, N.J., Bruggner, R.V., Rutler, R., et al. (2007). VectorBase: a home for invertebrate vectors of human pathogens. Nucleic Acids Res **35,** D503–D505.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25,** 955–964.

Lukashin, A.V., and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. **26,** 1107–1115.

Marchler-Bauer, A., Anderson, J.B., Cherukuri, P.F., Deweese-Scott, C., Geer, L.Y., Gwadz, M., et al. (2005). CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res **33,** D192–D196.

Markowitz, V.M., Szeto, E., Palaniappan, K., Grechkin, Y., Chu, K., Chen, I.M., et al. (2008). The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. Nucleic Acids Res **36,** D528–D533.

Paskin, N. (2005). The DOI® Handbook Version 4.2.0. From doi:10.1000/182.

Peterson, J.D., Umayam, L.D., Dickinson, T.M., Hickey, E.K., and White, O. (2001). The Comprehensive Microbial Resource. Nucleic Acids Res **29,** 123–125.

Salzberg, S.L., Delcher, A.L., Kasif, S., and White, O. (1998). Microbial gene identification using interpolated Markov models. Nucleic Acids Res **26,** 544–548.

Snyder, E.E., Kampanya, N., Lu, J., Nordbert, E.K., Karur, H.R., Shukala, M., et al. (2007). PATRIC: the VBI PathoSystems Resource Integration Center. Nucleic Acids Res **35,** D401–D406.

Stein, L. (2001). Genome annotation: from sequence to biology. Nat Rev Genet **2,** 493–503.

Sugawara, H., Ogasawara, O., Okubo, K., Gojobori, T., and Tateno, Y. (2008). DDBJ with new system and face. Nucleic Acids Res **36,** D22–D24.

Tatusov, R.L., Fedorova, N.D., Jackman, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., et al. (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4,** 41.

Tetko, I.V., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Fobo, G., et al. (2005). MIPS bacterial genomes functional annotation benchmark dataset. Bioinformatics **21,** 2520–2521.

Valencia, A. (2005). Automatic annotation of protein function. Curr Opin Struct Biol **15,** 267–274.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., et al. (2008). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res **36,** D13–D21.

Address requests for reprints to:
Samuel V. Angiuoli
Institute for Genome Sciences
University of Maryland School of Medicine
20 Penn Street
Baltimore, MD 21201

E-mail: sangiuoli@som.umaryland.edu