Standards for Annotation of Regulatory Sequences in ORegAnno (DRAFT)

Obi L. Griffith[1], Stephen Montgomery[2], and Steven Jones[1]

1. Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, Canada.
2. Welcome Trust Sanger Institute, Cambridge, UK.

NOTE: The author list and order above are not finalized. Additional contributors are expected.

I. Background

A large number of annotation efforts have been applied to genome sequence since the widespread adoption of genome wide sequencing efforts (reviewed recently by Brooksbank and Quackenbush (2006)[1]). Numerous tools exist to annotate genomes[2] and a large number of annotation efforts have been conducted or are underway. In general, these efforts have focused on coding sequences (e.g. defining gene models) and their function. Standards, rules or guidelines have been proposed for the annotation of nucleotide sequences[3], protein function (Swiss-Prot; the Gene Ontology)[4, 5], protein interactions (IntAct)[6], microarray experiments (MIAME)[7], biochemical models (MIRIAM)[8], biomedical text[9], phylogenetic analyses (MIAPA)[10], in situ hybridization and immunohistochemistry experiments (MISFISHIE)[11], and others. In some cases, functional non-coding sequences are included in these systems. But, to our knowledge, no guidelines or standards have been proposed specifically for the annotation of non-coding regulatory sequences or sequence variants.

II. Goal or purpose of annotation standards

The goal of this document is to outline the guidelines and minimal information required for annotation of a regulatory sequence for entry into the Open Regulatory Annotation (ORegAnno) system (www.oreganno.org)[12]. However we hope these standards will prove useful for other databases of regulatory sequences and for authors of experimental validation studies for newly discovered regulatory sequences. By adhering to the standards outlined below we hope to create a dataset of high quality that meets the following objectives:

1) To provide positive and negative control datasets for development and implementation of algorithms for the de novo identification of novel regulatory elements.
2) To provide training datasets for the development of text-mining tools for the automation of regulatory element annotation from the literature.
3) To provide a general purpose resource of known regulatory sequences for genetic and genomic studies of gene regulation.

NOTE: The following standards are in draft status. This document will be circulated and posted in a Wiki format for discussion, modification, and suggestions at the Regcreative meeting (http://www.dmbr.ugent.be/bioit/contents/regcreative/index.php). It is hoped that through discussion and contributions from a wide range of domain experts, a consensus will emerge and a final set of standards decided upon that addresses the needs of as wide a community as possible.

III. Annotation standards for an ORegAnno record

A. Minimal information for an ORegAnno record

1. Publication identifier

Each annotation must reference a valid PubMed article. This ensures that any record can be verified or validated by referring back to the original source. Before annotation, the publication must be entered into the ORegAnno publication queue by PMID. The paper will be assigned a status of ‘pending’ until the user begins annotation. Before

commencing annotation, the publication must be 'checked out' of the queue by setting its status to 'open'. Only the user who has opened that publication will then be allowed to annotate it. Upon completion of annotation the user should set the paper status to 'closed' and assign a reason for closure from one of the provided options. This process prevents users from creating redundant records or annotating the same publication concurrently. Unpublished results should not be entered. In the case of large datasets (e.g. obtained from other databases) a publication describing the source should still be provided. Also, an oreganno 'dataset' record should be created referencing the publication and providing a url for the database. Then, each record should include a reference to that dataset (see optional dataset id below).

Discussion item: Should a record reference only one publication? What should be done in cases where several papers describe experimental validation of the same regulatory sequence?

2. Regulatory sequence type

2.1. Transcription factor binding site (TFBS)

A transcription factor binding site record is a noncoding DNA sequence that is bound by a particular transcription factor to alter the expression of a particular gene. An example might be an experimentally confirmed Sp1 binding site.

2.2. Regulatory region

A regulatory region is a noncoding DNA sequence that is known to alter the expression of a particular gene. Canonical examples of regulatory regions are promoters and enhancers.

Discussion item: Should further sub-categorization of regulatory regions be allowed (e.g. Silencer, enhancer, locus-control region, etc)

2.3. Regulatory polymorphism

A regulatory polymorphism record is a noncoding DNA sequence that may or may not be bound by a known transcription factor in vivo, but has a variant that is confirmed to alter the expression of a particular gene. An example might be an experimentally confirmed Sp1 binding site that has two allelic variants, one of which, when present, downregulates its target gene relative to the other allele.

2.4. Regulatory haplotype

A regulatory haplotype record is a noncoding DNA sequence that contains many alleles in linkage disequilibrium (LD) that are confirmed to alter the expression of a particular gene. This is different than a regulatory polymorphism as the specific causal variant may not be known, only the alleles that are in LD with it.

3. Species

Each annotation must be attributed to a species which has a taxonomy id in the NCBI Taxonomy database[13]. The species under study should be explicitly stated in the source publication. However, if it is not, the species can be inferred with caution by the gene

identifier or sequence. In particular, annotators should be wary of regulatory sequences from one species that are assayed for function in another species (e.g. a human sequence might be assayed for function in a mouse model). In such cases, the sequence should be annotated for the species from which the sequence was derived.

Discussion item: In a case where sequence conservation is perfect between the species of interest and model organism (e.g. both mouse and human have identical regulatory sequence upstream of an orthologous gene) could an assay in one be considered evidence for function of the sequence in both species?

4. Target Gene

Each annotation describes a regulatory property of one target gene which should be identified by Entrez Gene ID, EnsEMBL ID or a user-defined identifier.

5. Transcription factor

Each transcription factor binding site or regulatory mutation must specify a target transcription factor which is either user-defined, in Entrez Gene or in EnsEMBL. If there is no recorded gene target, a classification of "unknown" is allowed.

Discussion item: Should multiple TFs be allowed for a single record? Or should this form a second record?

Discussion item: Should TF complexes be allowed?

6. Sequence and flanking sequence

Each annotation should include the functional or bound sequence as reported in the publication. To facilitate mapping to current genome coordinates, sufficient flanking sequence should be provided from the most current genome build. Specifically, each record should include at least 40 bases (ideally ~100 bases) of flanking genomic sequence. If the bound sequence is different from the reference genome sequence the record should not be entered unless the differences can be verified as known polymorphisms. Before final entry of sequence and flank a check against the current reference genome with a sequence alignment program (such as BLAT or BLAST) should be performed to confirm that it aligns unambiguously and with no unexpected base discrepancies. For regulatory mutations, each variant that has been proven to cause a change in gene expression is a separate record. The sequences containing both the wild-type and mutant sequences must be specified. For clarity, the bound sequence or regulatory polymorphism sequence should be entered as upper case and the flanking sequence as lower case (e.g. atcgtacgtaCGCGGGCattcgacat). This is particularly important if the binding site is present in more than one instance in the flanking sequence.

7. Experimental evidence

Each annotation specifies an evidence type, subtype and class describing the biological technique cited to discover the regulatory sequence. Each annotation can have multiple entries from any evidence class, type and subtype describing each piece of experimental evidence for the regulatory sequence and/or binding protein. As a minimum, a record

must have at least one piece of *in vivo* or *in vitro* experimental evidence to be considered suitable for entry into ORegAnno. *In silico* or indirect evidence (e.g. evidence type: 'Sequence conservation') should be entered as supplemental evidence only.

Discussion item: What is the minimal evidence we should allow?

7.1. Evidence Class

Evidence classes are broken into two categories: the 'regulator' classes describe evidence for the specific protein(s) that bind a site. The 'regulatory site' classes describe evidence for the function of a regulatory sequence itself. These two categories are further divided into three levels of regulation (transcription, transcript stability and translation). Thus, a total of six evidence classes currently exist.

Discussion item: Are there other evidence classes that should be included?

7.2. Evidence type and subtype

Evidence types describe the generic assay used while subtypes define specific implementations of these assays. For regulatory polymorphisms or haplotypes, association studies (evidence type: 'Association study') alone should not be considered sufficient evidence as these studies typically can not distinguish a functional polymorphism from a non-functional polymorphism in linkage disequilibrium with the functional polymorphism. The evidence type 'Literature derived' should only be used in cases where sequences were manually curated by another group of experts adhering to standards materially equivalent to those outlined in this document but where specific experiments were not recorded or can not be confidently mapped to the evidence ontology. If no evidence type or subtype exists to describe the experimental evidence reported, a new evidence type or subtype should be submitted for inclusion in the evidence ontology with a detailed description and example of the method.

Discussion item: Should ORegAnno migrate to a more complex or formal ontology system for evidence.

7.3. Cell type

In many cases, the functional validation of a regulatory sequence depends on the context under which it was assayed. One important factor determining this context is the cell-type. Therefore, wherever possible, the cell-type in which experiments were conducted should be recorded for each piece of experimental evidence. If a particular experiment (e.g. a reporter gene assay) is performed in several different cell types (e.g. different cell lines) these can be considered multiple pieces of evidence (one for each cell type). ORegAnno currently uses the eVOC cell-type ontology for this purpose[14].

7.4. Evidence comment

An evidence comment should be provided to describe in detail the specific implementation and results of the experiment for each line of evidence supporting the record.

8. Outcome

Each record is associated to a positive, neutral or negative outcome based on the experimental results from the primary reference. For instance, a sequence that was demonstrated not to bind a particular transcription factor could be annotated as a negative outcome; however, to be meaningful, the associated evidence must provide adequate information to determine the conditions assayed. In general, records labeled with a negative outcome require a higher burden of proof than those with positive because a lack of activity under one condition (cell type, developmental stage, etc) does not imply lack of function under any other condition. In order to create a useful 'negative control set', the 'negative outcome' status should be limited to those sequences which have been tested for large range of conditions. Records with a 'neutral outcome' have uncertain value. However, a user may wish to include such records for completeness when comprehensively annotating a paper which tested multiple regulatory sequences with different outcomes.

9. User information

In order to create a sense of accountability and ownership for data contributed to the community, all records are associated with the user who created them. Each user is associated with the name, affiliation and email address entered upon creation of their account. Additionally, users belong to one of three roles: user, validator and administrator. A user role enables a contributor to add individual annotations to the database. A validator role enables a contributor to score individual annotations in the database. Validators can modify the overall score for an annotation based on their ability to confirm the accuracy of the annotation from literature. An administrator role enables a contributor to assign roles, add or define evidence (classes, types, and subtypes) and batch upload large sets of annotations directly to the database. Both administrator and validator roles allow the modification of records; for a record modification, a new record is created and the old record is marked as being deprecated by the newer record. Each role is further permitted to add comments to individual annotations to improve subsequent users' understanding of a particular annotation.

B. Optional information for an ORegAnno record

1. Dataset id

Each annotation can optionally be associated to a specific dataset. This functionality allows external curators to manage particular sets of annotation using ORegAnno's curation tools. Before a record can be associated with a dataset, an ORegAnno dataset record must be created. This record should include a description of the dataset, a url for the data source, and citation for the original publication describing the dataset.

2. Locus name

A regulatory sequence can have a locus name associated to it. This can be user defined or even refer to another ORegAnno record.

3. Sequence search space

Where available, any annotation can provide search space information specifying the region that was assayed, not just the regulatory sequence. This could be used in cases

where a series of promoter deletions and assays were performed to identify the minimal promoter required for maximal transcription activity.

4. Optional information for regulatory polymorphism records

4.1. Variant sequence and identifier

If available, a dbSNP ID can be specified for a regulatory polymorphism record.

4.2. Type of variant

The type of variant is specified as either being germline, somatic or artificial. If the variant corresponds to a known polymorphism it is most likely a germline variant. However, unless sequence comparisons of parent/offspring, tumor/normal, etc are performed it can be difficult to distinguish between somatic and germline mutations. If the variant was created solely through experimental manipulation (e.g. site-directed mutagenesis) and was not observed in any individual or organism it should be entered as 'artificial'. If the type can not be ascertained it should be entered as 'uncertain'.

5. Meta-data

Each annotation in ORegAnno is optionally associated one or more administrator-defined meta-data types. These can be additional pieces of data relevant to a specific type of annotation or those captured in important datasets that would be useful to append to an ORegAnno annotation. Each added meta-data element must match a pattern defined by the administrator for these meta-data types. The value of the meta-data element is also inserted into a meta-data type specific URL for cross-referencing to other browsers

6. Comment

Each record can be accompanied by one or more comments. These can be useful for explaining specific problems with annotation or for important conclusions or data for which no appropriate field exists in the database. Any ORegAnno user can add additional comments to a record.

IV. Advice for authors of experimental studies

To encourage researchers to report their experimental validations of regulatory sequences in a form with maximal utility to the community we offer the following suggestions. The species under study should be explicitly stated. Particularly in cases where model organisms are used to investigate human sequences it should be made clear whether it is the human sequence or orthologous model organism's sequence being assayed. Always provide a stable, unique identifier such as Entrez or Ensembl gene ID for the target gene of interest and transcription factor. If one is not yet available, a sequence identifier (such as Refseq or genbank accession) should be provided for the gene or transcript sequence. Always provide the actual sequence in question and sufficient flanking sequencing to unambiguously map to genomic coordinates. In lieu of the actual sequence, sequence coordinates for a specific genome assembly should be provided. Never identify a sequence by relative coordinates only. A common practice is to refer to a sequence relative to a transcription start site (TSS). However, this assumes that the reader knows exactly which transcript is being referenced. Many genes will have multiple TSSs. TSSs are frequently revised and upstream sequences may change in subsequent genome

assemblies. Finally, experimentalists are encouraged to add their own findings into ORegAnno as they are obviously the most qualified to do so.

References

1. Brooksbank, C. and J. Quackenbush, *Data standards: a call to action.* Omics, 2006. **10**(2): p. 94-9.
2. Rouze, P., N. Pavy, and S. Rombauts, *Genome annotation: which tools do we have for it?* Curr Opin Plant Biol, 1999. **2**(2): p. 90-5.
3. Cochrane, G., et al., *Evidence standards in experimental and inferential INSDC Third Party Annotation data.* Omics, 2006. **10**(2): p. 105-13.
4. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Res, 2003. **31**(1): p. 365-70.
5. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
6. Hermjakob, H., et al., *IntAct: an open source molecular interaction database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D452-5.
7. Brazma, A., et al., *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.* Nat Genet, 2001. **29**(4): p. 365-71.
8. Le Novere, N., et al., *Minimum information requested in the annotation of biochemical models (MIRIAM).* Nat Biotechnol, 2005. **23**(12): p. 1509-15.
9. Wilbur, W.J., A. Rzhetsky, and H. Shatkay, *New directions in biomedical text annotation: definitions, guidelines and corpus construction.* BMC Bioinformatics, 2006. **7**: p. 356.
10. Leebens-Mack, J., et al., *Taking the first steps towards a standard for reporting on phylogenies: Minimum Information About a Phylogenetic Analysis (MIAPA).* Omics, 2006. **10**(2): p. 231-7.
11. Deutsch, E.W., et al., *Development of the Minimum Information Specification for In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE).* Omics, 2006. **10**(2): p. 205-8.
12. Montgomery, S.B., et al., *ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation.* Bioinformatics, 2006. **22**(5): p. 637-40.
13. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2000. **28**(1): p. 10-4.
14. Kelso, J., et al., *eVOC: a controlled vocabulary for unifying gene expression data.* Genome Res, 2003. **13**(6A): p. 1222-30.