ORegAnno: Open Regulatory Annotation (www.oreganno.org)

An open access database and curation system for regulatory sequences

Griffith OL[1,2], Montgomery SB[1,2], Sleumer MC[2], Bergman, CM[3], Bilenky M[2], Pleasance ED[2], Prychyna Y[2], Zhang X[2], Jones SJM[2]

1. These authors contributed equally to this work; 2. Canada's Michael Smith Genome Sciences Centre; 3. University of Manchester

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

Canada's Michael Smith Genome Sciences Centre
www.bcgsc.ca

1. Abstract

Our understanding of gene regulation is currently limited by our ability to collectively synthesize and catalogue transcriptional regulatory elements stored in scientific literature. Over the past decade, this task has become increasingly challenging as the accrual of biologically-validated regulatory sequences has accelerated. Here, we present the Open Regulatory Annotation (ORegAnno) database as a dynamic collection of literature-curated regulatory regions (promoters, enhancers, etc), transcription factor binding sites, and regulatory mutations (SNPs and haplotypes). ORegAnno is a web resource that has been designed to manage the submission, indexing, and validation of new annotations from users worldwide. Submissions to ORegAnno are immediately cross-referenced to EnsEMBL, dbSNP, Entrez Gene, the NCBI Taxonomy database, and PubMed, where appropriate. ORegAnno currently contains 1804 binding sites, 780 regulatory regions, and 107 regulatory polymorphisms or haplotypes from 9 species. We are currently in the process of adding a large number of additional records from the literature and species-specific databases. The ORegAnno resource represents the first open-access community-based forum for annotation of cis-regulatory sequences. It is also the first system to incorporate structured experimental evidence and allow both negative and positive results. The requirements for sufficient flanking sequence and verified gene identifiers (Ensembl or Entrez) ensure maximum compatibility with the community's various research needs. This set of experimentally verified regulatory sequences represents a valuable resource for researchers investigating transcriptional regulation or regulatory variation and provides an open-access system for continued, community based accumulation of sites within a standardized framework. It also forms an integral part in the evaluation of our own cis-regulatory element predictions (www.cisred.org). For convenience, ORegAnno is available directly through MySQL, Web services, and online at www.oreganno.org.

Transcription Factor Binding Site Resources:

> TRANSFAC
www.biobase.de

> Drosophila DNase I Footprint Database
www.flyreg.org

> Transcription Regulatory Regions Database (TRRD)
www.bionet.nsc.ru/trrd/

> Transcriptional Regulatory Element Database (TRED)
rulai.cshl.edu/TRED/

> Riken Transcription Factor Database (TFDB)
genome.gsc.riken.jp/TFdb/

> plantCARE
intra.psb.ugent.be:8080/PlantCARE/

> Arabidopsis thaliana Promoter Binding Element Database (AtProbe)
rulai.cshl.edu/cgi-bin/atprobe/atprobe.pl

> The Arabidopsis cis-regulatory element database (AtcisDB)
arabidopsis.med.ohio-state.edu/AtcisDB/

Regulatory Region Resources:

> Hematopoiesis Promoter Database (HemoPDB)
bioinformatics.med.ohio-state.edu/HemoPDB/

> MPromDb Mammalian Promoter Database
rulai.cshl.edu/CSH_mpdb/

> Osteo - Promoter Database (OPD)
www.cpd.rau.ac.il/

> Orthologous Mammalian Gene Promoter database (OMGProm)
bioinformatics.med.ohio-state.edu/OMGProm/

> Arabidopsis transcription factor database (AtTFDB)
arabidopsis.med.ohio-state.edu/AtTFDB/

> Eukaryotic Promoter Database (EPD)
www.epd.isb-sib.ch/

> PlantProm DB
mendel.cs.rhul.ac.uk/mendel.php

> Promoter Database of Saccharomyces cerevisiae (SCPD)
rulai.cshl.edu/SCPD/

> C. elegans promoter database (CEPDB)
rulai.cshl.edu/cgi-bin/CEPDB/home.cgi

> The Liver Specific Gene Promoter Database
rulai.cshl.edu/LSPD/

Regulatory Variant Resources:

> rSNP_Guide
www.ngs.bionet.nsc.ru/rsnpg/systems/rsnp/

> Human Gene Mutation Database (HGMD)
www.hgmd.org/

> dbQSNP
gene.gen.kyushu-u.ac.jp/

> PromoSign
poly.wustl.edu/promsig/main.html

2. Design

Figure 1. The ORegAnno Resource

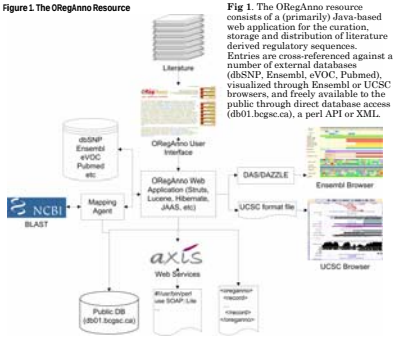

Fig 1. The ORegAnno resource consists of a (primarily) Java-based web application for the curation, storage and distribution of literature derived regulatory sequences. Entries are cross-referenced against a number of external databases (dbSNP, Ensembl, eVOC, Pubmed), visualized through Ensembl or UCSC browsers, and freely available to the public through direct database access (db01.bcgsc.ca), a perl API or XML.

Figure 2. Database schema for mySQL

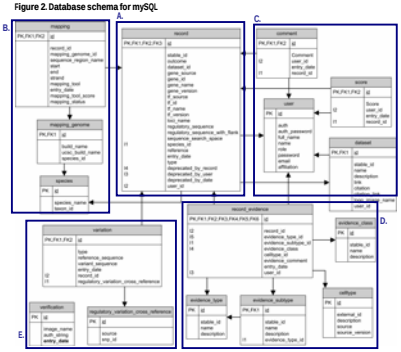

Fig 2. (A) Every ORegAnno record consists of a stable id, record type, species, reference, outcome, target gene, transcription factor (if known), sequence and flank; (B) The sequence and species are used to derive genomic coordinates by BLAST alignment; (C) Each record is associated with the user who entered it as well as the history of comments and scores it has received. If the record was acquired from an existing database it will be linked to that dataset's information; (D) If the record is a polymorphism or haplotype the variant sequence is also stored as well as any external links for that variant; (E) Each record will normally have some evidence for the function of the sequence from the original publication. This evidence is categorized according to several classes, types, and subtypes (see table 2). If known, the cell type used for the experiments can also be stored using the eVOC cell type ontology[1].

3. Implementation

Figure 3. The ORegAnno User Interface

Fig 3. The ORegAnno user interface provides: (A) Login status; (B) Current contents of database (with link for detailed view); (C) Options to login/logout or create a new user (login only required for annotation); (D) Search engine (powered by Lucene) for basic or advanced searching; (E) Annotation forms for regulatory regions, binding sites, polymorphisms or haplotypes (login required); (F) Tools for locating regulatory sites by sequence or position for an Ensembl or Entrez target gene; (G) Database downloads/access are available through regular xml dumps, direct mysql database access, or a perl API (using SOAP); (H) Help documentation provides walkthroughs, guidelines for annotation, and other useful information; (I) A citation page gives credit to major contributors and links to a complete ORegAnno user list.

Figure 4. An ORegAnno record

Fig 4. For each record in ORegAnno: (A) a stable, unique identifier is assigned; (B) Detailed views of comments, score history, and evidence are available; (C) A record can be one of four types (Transcription factor (TF) binding site, regulatory region, regulatory polymorphism, or regulatory haplotype). Outcome indicates if experiments proved or disproved a functional role for the sequence. Ensembl or NCBI Entrez Gene IDs are provided for both the target gene and TF (if available). Each record must also include a taxon ID, pubmed ID, target sequence and sufficient flank for genome alignment. (D) User information is available (email, user name, full name, and affiliation). A user can belong to one of three roles (user, validator, or administrator); (E) evidence for the record is documented according to ORegAnno evidence types (see table XXX for examples); (F) Validators can validate a record or invalidate a record by giving it a positive or negative score; (G) Sequences are automatically mapped to genome coordinates and can be viewed in UCSC or Ensembl genome browsers.

4. Evidence

Table1. Sample of Evidence types and subtypes

Evidence type	Evidence subtype
Electrophoretic Mobility Shift Assay (EMSA)	Direct gel shift
	Supershift
	Gel shift competition
Reporter Gene Assay	Transient transfection luciferase assay
	Chloramphenicol acetyltransferase (CAT) Assay
	In-vivo GFP Expression Assay
	Dual luciferase reporter gene assay
	In-vivo LacZ Expression Assay
Protein Binding Assay	Chromatin immunoprecipitation (ChIP)
	DNase Footprinting Assay
	Yeast 1-hybrid assay

Table1. Each ORegAnno record is associated with one or more pieces of evidence. Oreganno currently contains 9 types and 30 subtypes of evidence. A user with administrator status can add new evidence types and subtypes as needed.

5. Database contents

Table 2. Current contents of ORegAnno database

Species	Regulatory Haplotype	Regulatory Polymorphism	Regulatory Region	Transcription Factor Binding Site
Caenorhabditis briggsae	0	0	0	24
Caenorhabditis elegans	0	0	8	117
Danio rerio	0	0	2	0
Drosophila melanogaster	0	0	0	1331
Gallus gallus	0	0	0	13
Homo sapiens	4	103	765	196
Mus musculus	0	0	1	87
Rattus norvegicus	0	0	4	35
Xenopus tropicalis	0	0	0	1
Totals	**4**	**103**	**780**	**1804**

Table 2. ORegAnno currently contains 2691 entries from 20 users. These include 780 regulatory regions, 1804 transcription factor binding sites, and 107 regulatory mutations (polymorphisms and haplotypes) from 9 species. A large fraction of these sites were obtained from previous large-scale collections such as the FLYREG resource [2] and a large set of muscle/liver-specific regulatory sites curated by Wasserman and Fickett [3,4]. 11 regulatory polymorphism records were obtained from rSNP_DB [5]; rSNP_DB records were filtered to include only those records which pertained to natural mutations or polymorphisms. In addition, over 200 new annotations were obtained by manual curation of literature.

6. Visualizations

Figure 5. Genome browser views for ORegAnno records

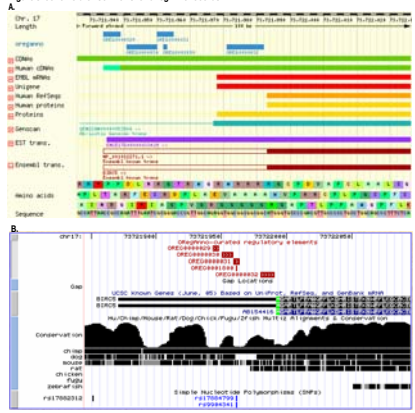

Fig 5. (A) Ensembl and (B) UCSC views allow the user to visualize any ORegAnno sequence in its genomic context.

7. Conclusions

> A large collection of functionally-validated regulatory annotations available with unrestricted access.
> An open-access system for community based accumulation of sites within a standardized framework.
> Incorporates a structured system for experimental evidence.
> A useful resource for computational investigations of gene regulation.

8. Acknowledgments

We would like to acknowledge the Wasserman lab (http://www.cisreg.ca/tjkwon/), and James Fickett (http://www.cbil.upenn.edu/MTIR/HomePage.html) for generously making their regulatory element catalogues publicly available. We thank the ORegAnno users for their continuing efforts to improve this resource through manual curation and record validation.

funding } We gratefully acknowledge funding from Genome Canada, Genome British Columbia and the BC Cancer Foundation. SBM was supported by the Natural Sciences and Engineering Research Council (NSERC) and the Michael Smith Foundation for Health Research (MSFHR). OLG was supported by the Canadian Institutes of Health Research (CIHR), NSERC and MSFHR. EDP was supported by CIHR. MCS and SJMJ were supported by MSFHR.

references } 1. Kelso *et al.* 2003; 2. Bergman *et al.* 2005; 3. Ho Sui *et al.* 2005; 4. Wasserman and Ficket. 1998; 5. Ponomarenko *et al.* 2001.