

ORegAnno: A Community-Based Annotation System for Literature-Derived Regulatory sequences

Griffith OL^{1,†}, Montgomery SB^{1,†}, Bergman CM², Bilenky M¹, Chu B¹, Pleasance ED¹, Prychyna Y¹, Sleumer MC¹, Zhang X¹ and Jones SJM¹

¹ Canada's Michael Smith Genome Sciences Centre, Vancouver, Canada

² University of Manchester, Manchester, UK

[†] These authors contributed equally to this work

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 2084 binding sites, 1554 regulatory regions, and 181 regulatory mutations from 13 species. A publication queue allows submission of papers of interest to the community for annotation and currently houses over 4000 papers related to gene regulation. 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 a structured experimental evidence ontology 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 and open-source system for continued, community based accumulation of sites within a standardized framework. It also forms an integral part in the evaluation of our own de novo cis-regulatory element predictions (www.cisred.org). ORegAnno is available directly through MySQL, Web services, the UCSC genome browser, and online at www.oreganno.org.