ORegAnno 3.0: A community-driven resource for curated regulatory annotation

Robert Lesurf
Washington University School of Medicine in St. Louis

Kelsy C. Cotto
Washington University School of Medicine in St. Louis

Grace Wang
Washington University School of Medicine in St. Louis

Malachi Griffith
Washington University School of Medicine in St. Louis

Katayoon Kasaian
BC Cancer Agency

See next page for additional authors

Follow this and additional works at: http://digitalcommons.wustl.edu/open_access_pubs

Recommended Citation

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact engeszer@wustl.edu.
Authors
Robert Lesurf, Kelsy C. Cotto, Grace Wang, Malachi Griffith, Katayoon Kasaian, Steven J M Jones, Stephen B. Montgomery, Obi L. Griffith, and The Open Regulatory Annotation Consortium

This open access publication is available at Digital Commons@Becker: http://digitalcommons.wustl.edu/open_access_pubs/4395
ORegAnno 3.0: a community-driven resource for curated regulatory annotation

Robert Lesurf1, Kelsy C. Cotto1, Grace Wang1, Malachi Griffith1,2,3, Katayoon Kasaian4, Steven J. M. Jones4,5,6, Stephen B. Montgomery7,8,* Obi L. Griffith1,2,3,9,* and The Open Regulatory Annotation Consortium

1McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, USA, 2Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA, 3Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO 63110, USA, 4Canada’s Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, BC V5Z 4S6, Canada, 5Department of Molecular Biology & Biochemistry, Simon Fraser University, Burnaby, BC V5A 1S6, Canada, 6Department of Medical Genetics, University of British Columbia, Vancouver, BC V6T 1Z3, Canada, 7Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA, 8Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA and 9Department of Medicine, Division of Oncology, Washington University School of Medicine, St. Louis, MO 63110, USA.

Received September 16, 2015; Revised October 24, 2015; Accepted October 26, 2015

ABSTRACT

The Open Regulatory Annotation database (ORegAnno) is a resource for curated regulatory annotation. It contains information about regulatory regions, transcription factor binding sites, RNA binding sites, regulatory variants, haplotypes, and other regulatory elements. ORegAnno differentiates itself from other regulatory resources by facilitating crowd-sourced interpretation and annotation of regulatory observations from the literature and highly curated resources. It contains a comprehensive annotation scheme that aims to describe both the elements and outcomes of regulatory events. Moreover, ORegAnno assembles these disparate data sources and annotations into a single, high quality catalogue of curated regulatory information. The current release is an update of the database previously featured in the NAR Database Issue, and now contains 1,948,307 records, across 18 species, with a combined coverage of 334,215,080 bp. Complete records, annotation, and other associated data are available for browsing and download at http://www.oreganno.org/.

INTRODUCTION

The Open Regulatory Annotation database (ORegAnno) was first released about a decade ago (1), with the intention to collect and synthesize a catalogue of regulatory elements. It remains unique in the field because of its focus on collecting high quality, curated regulatory records from the literature. Moreover, ORegAnno relies on a thriving community of scientists who are interested in contributing to the resource, as well as utilizing its data. Since the last release of ORegAnno in early 2008 (2), the amount and types of published regulatory data have grown exponentially. This relates in part to high-throughput studies from the ENCODE consortium and others, who have performed an enormous number of ChIP-seq, DNase-seq, FAIRE-seq and other experiments aiming to identify biochemically available and transcriptionally active regions of genomes (3). While these efforts are excellent resources for identifying candidate regulatory regions, ENCODE efforts have suggested that as much as 80% of the genome could be functional (3). This controversial finding has been the focus of much attention in the community, with several commentaries pointing out that these types of high-throughput data are prone to overestimates due to experimental and statistical methods that result in a high number of false positive calls (4–6). Moreover, they do not necessarily provide a comprehensive understanding of all of the elements involved in gene regulation. For example, knowing the region of DNA that is bound by a transcription factor does not directly indicate whether the expression of any genes are altered, nor whether an alteration results in up- versus down-regulation. Validation of the genomic regions identified by ENCODE and others requires a large number of low-throughput experimental data paired with careful manual curation. Additionally, much of the available evidence supporting gene regu-
loration is dispersed across various experiments, specialized datasets, and individual publications, making it cumbersome to obtain regulatory information that has been released by the community across this broad set of sources. The current version of ORegAnno seeks to address these issues by cataloging a large number of new, curated, high-quality regional records that are derived from published literature and other data resources.

RESULTS

Overview

The current version of ORegAnno now has a total of 1,948,307 unique records. These records cover a combined 334,215,080 bp across 18 species (Figure 1A and B). The vast majority of these records are mapped to human and mouse genomes, with 1,452,466 records in human (261,660,516 bp in the GRCh38/hg38 genome assembly version) and 415,808 records in mouse (57,253,973 bp in the GRCm38/mm10 genome assembly version).

As a measure of the success of our community-based participation, ORegAnno currently has 1,044 registered users. Aside from the principal authors of this paper, 13,301 records have been contributed by members of the broader community (The Open Regulatory Annotation Consortium). ORegAnno continues to have a robust verification system to ensure that contributed records are accurate and appropriately annotated. A set of trusted consortium members have been granted a ‘validator’ status, allowing them to review and up- or down-vote records. This results in individual record scores that are visible to all users. Moreover, when a record is negatively scored, it will typically be assigned a deprecated status. ORegAnno additionally includes an ontology for summarizing the experimental evidence that supports the regulatory elements and outcome in each record. Together, these features allow users to filter records according to various quality criteria.

The ORegAnno database has served as a repository for publishing regulatory sites derived from experimental data (7), and it has been incorporated into other resources including the Babelomics (8), cisRED (9), ConTra (10), GRASP (11), i-cisTarget (12), LASAGNA-Search 2.0 (13), the UCSC Genome Browser (14) and more. Similarly, the annotated information included in ORegAnno has been used to construct gene regulation networks for the development of other tools and the analysis of gene expression data (15–19). ORegAnno records were used in the REC-set design for a capture sequence reagent (20), and as part of the definition for regulatory sites of the human genome (tier 2) in the Genome Modeling System (21), an analysis information management system at the McDonnell Genome Institute of Washington University that has been used to process over 4,800 human whole genome samples, over 40,000 exomes, and over 1,400 transcriptomes. Similarly, ORegAnno has been adapted into the information systems of other research centers including the Broad Institute and Cancer Research UK, where it has been used in the analysis of several high impact studies (22–25).

Because ORegAnno focuses on curated regulatory information, the total genomic coverage found in ORegAnno is smaller than that identified by resources such as ENCODE or the ENSEMBL regulatory tracks (26), which are largely a summary of ENCODE data (Figure 2). This trade-off is part of an effort to ensure that ORegAnno represents a high-quality curated set of regulatory elements, with the aim of maintaining a low number of false positive records.

Updates

Older records, including those that were added through crowd-sourcing efforts via the web, have been updated to ensure that only accurate and up-to-date gene symbols are being used. This was accomplished through a combination of automatically updating symbols using NCBI Gene or ENSEMBL identifiers, as well as by manually checking incorrect and missing data. In addition, previously missing identifiers from NCBI Gene or ENSEMBL have been added where possible, allowing for future automated updates to ensure the accuracy of these gene lists. These updates have resulted in 423 automated changes and 13,174 manually curated changes (13,597 total) affecting 10,386 records.

For all ORegAnno records (existing and new), genomic coordinates have been updated and expanded using liftOver (27). This involved converting older genomic coordinates to newer assembly versions, as well as converting coordinates from new versions to older assemblies. Thus, each record may now be associated with multiple genomic coordinates (from multiple assembly versions). For example, since the last version of ORegAnno was published in 2008, the human genome assembly version GRCh38/hg38 was released. All existing ORegAnno human records having genomic coordinates based on assembly versions GRCh36/hg18 or GRCh37/hg19 now have additional updated coordinates using GRCh38/hg38. Similarly, new records that were entered using GRCh38/hg38 coordinates have received additional coordinates based on GRCh37/hg19 and GRCh36/hg18. This allows users to access the genomic coordinates of regulatory regions for the assembly versions that best suit their purposes. Finally, new types of transcriptional regulation have been defined in the current release (Figure 1C and D). These includes microRNA and small non-coding RNA binding sites, as well as operons that function to regulate multiple genes under a single promoter.

New records

ORegAnno has maintained a focus on incorporating records derived from high quality, manually curated evidence for gene regulation. These typically include experimental evidence demonstrating that binding of a regulatory element to a specific region of DNA or RNA alters corresponding gene expression levels. In total, the current release of ORegAnno contains 2010 unique records covering 112,582 bp derived directly through literature curation, including 661 records that have been added since the previous ORegAnno release.

Highly validated external databases that had been incorporated into earlier ORegAnno releases have been updated. This includes 1,874 new records covering an additional 3,591,656 bp derived from VISTA Enhancers (28) (2,196 total records covering 3,966,796 total bp), 2,934 new records covering an additional 863,201 bp derived from the Yeast
Figure 1. Current content of the ORegAnno database. Content statistics are divided by species (A and B), regulatory type (C and D), and data source (E and F).
Figure 2. Comparison of the genomic coverage captured by ORegAnno and the ENSEMBL Regulatory Track. A Venn diagram demonstrates coverage overlaps for human genome assembly version GRCh38/hg38, with sets sized to scale. The ENSEMBL Regulatory Track is divided into two main sets, a track overview set and the transcription factor binding site (TFBS) set.

Regulatory Map (29) (7320 total records covering 899,449 total bp), as well as 2051 new transcription factor binding site records covering an additional 29,405 bp derived from REDfly (30) (2695 total records covering 913,486 total bp). Previously, ORegAnno had imported records from FlyReg (31), which has since been merged into REDfly.

New records have been created by importing data from external databases that were not found in previous ORegAnno releases. This includes 1,093,443 records covering 11,780,604 bp imported from the JASPAR CORE database (32), which contains a curated, non-redundant set of experimentally obtained transcription factor binding sites in eukaryotes. 783,742 records covering 300,003,052 bp were imported from the PAZAR database (33), which included only records with curated evidence of transcription factor binding and regulatory sequence annotation across various species. 11,451 records covering 4,194,677 bp were derived from RegulonDB (34), a database of transcriptional regulation in Escherichia coli K-12, and includes manually curated records that have been complemented with high-throughput datasets and comprehensive computational predictions. We combined conserved miRNA target site predictions from miRanda-mirSVR (35) with experimentally-validated miRNA-target interaction data from miRTarBase (36), leading to the addition of 3,072 new ORegAnno records covering 44,353 bp. 131 records covering 1216 bp were derived from NFI-RegulomeDB (37), a database with curated binding sites for the NFI (Nuclear Factor I) family of transcription factors using data from the published literature. Finally, 51 transcription factor binding site records covering 7503 bp were created from the PCNE database of phylogenetically conserved noncoding elements (38).

Because of the open and accessible design of the ORegAnno database and website, ORegAnno has been used for submitting published experimental data. Since the previous ORegAnno release, four datasets derived from high throughput studies have been submitted, and were subsequently curated to ensure that only regulatory regions with a high degree of evidence were retained. These include RELA (p65) ChIP–PET binding sites in human monocytes (39) (489 records covering 52,886 bp), ESR1 binding sites in human MCF-7 breast cancer cells (40) (1234 records covering 165,538 bp), Esr1 binding sites in mouse liver (41) (5568 records covering 2,378,460 bp), and Foxa2 binding sites in mouse liver (7) (11,475 records covering 8,236,933 bp). In all of these cases, DNA sequences were filtered according to signal strength and proximity to signal peak to reduce false positive calls. A summary of the number of records and genomic coverage contributed by each data source is shown in Figure 1E, F and Supplementary Table S1.

Data access

The ORegAnno database continues to be accessible under an open-source license (GNU Lesser General Public License), in order to encourage development and participation from the community. Monthly ORegAnno database summaries are automatically performed and provide fundamental regulatory information from ORegAnno in a tab-delimited text file that is available for free download. The OReganno website is available at https://java.net/projects/oreganno/.

The regulatory regions and associated annotation for all supported species have been submitted to the UCSC Genome Browser (14) as updates to existing ORegAnno tracks. This updates existing tracks with a more comprehensive collection of putatively regulatory elements, and additionally provides new tracks on several genome assembly versions.

Applications

Recently, there has been immense focus on the role of regulatory regions in cancer. In particular, recurrent somatic mutations in the TERT promoter have been identified in various cancer types (42–45), and are associated with increased expression of TERT. Although the importance of TERT up-regulation in cancer has been well-established for nearly two decades (46), it is only in recent years that we’ve identified the regulatory mechanism driving TERT up-regulation in such cases. While additional efforts have identified a small number of other recurrent regulatory mutations in cancer (47–49), this number is far smaller than the recurrent protein-coding mutations that have been identified. This is likely due to several factors, including that most
Figure 3. Capture reagent using ORegAnno sites improves coverage of regulatory regions in human hepatocellular carcinoma. (A) Coverage across the entire ‘regulome’ is visualized as a heatmap for each of the ten HCC cases. WRS samples have greater sequencing read depth across the targeted region, compared with WGS samples. (B) An illustrative IGV (51) screenshot is shown of the TERT promoter for one HCC case. A canonical C228T somatic mutation is observed in the WRS data, but cannot be reliably called in the WGS data.
cancer survey projects have focused primarily on coding regions by using exome capture reagents to enrich for these regions, and that the TERT promoter region, as with many other genes, has a high GC content making both PCR amplification and sequencing challenging.

Previous identification of coding regions of the genome made it possible to perform exome targeted sequencing of these regions in a large number of cancer cases at a relatively low cost. Similarly, we’ve used ORegAnno and other sources to design a ‘regulome’ capture reagent for targeted sequencing. This high quality, relatively small coverage of literature-curated transcription factor binding sites, regulatory polymorphisms, and NFI-RegulomeDB (37) sites identified in ORegAnno, in conjunction with regulatory regions defined by FunSeq (50), and 500 bp regions upstream of each gene transcription start site, were used to define the ‘regulome’ region. As a proof of principle, we then applied ‘regulome’ capture-sequencing to ten normal/tumor pairs of hepatocellular carcinoma (HCC). Overall coverage of the regulome region defined in the capture reagent was higher in whole regulome sequencing (WRS) samples versus whole genome sequencing (WGS) samples of the same tissues, with median average read depths of 29× in WGS normal, 49× in WGS tumor, 60× in WRS normal and 68× in WRS tumor (Figure 3A, Supplementary Table S2). This improved coverage allowed us to reliably identify the canonical somatic TERT promoter mutation C228T in six of the ten cases, an illustrative example of which is shown in Figure 3B.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
We thank the Open Regulatory Annotation Consortium for their continuing contributions to ORegAnno through the identification of relevant publications and manual curation efforts. We additionally thank the publishers of experimental data and external regulatory databases that were included in the current release of ORegAnno.

FUNDING
American Cancer Society [IRG-58-010-58 to O.L.G.]; Edward Mallinckrodt, Jr. Foundation (to S.B.M.); National Institutes of Health [K99HG007940 to M.G., R01HG008150 and R01MH101814 to S.B.M., K22CA188163 to O.L.G.]. Funding for open access charge: Departmental funding (to O.L.G.).

Conflict of interest statement. None declared.

REFERENCES


