SnoVault and encodeD: A novel object-based storage system and applications to ENCODE metadata

Benjamin C. Hitz[1] (hitz@stanford.edu)
Laurence D. Rowe[1], Nikhil R. Podduturi[1], David I. Glick[1], Ulugbek K. Baymuradov[1], Venkat S. Malladi[3], Esther T. Chan[1], Jean M. Davidson[1], Idan Gabdank[1], Aditi K. Narayana[1], Kathrina C. Onate[1], Marcus C. Ho[1], Brian T. Lee[2], Stuart R. Miyasato[1], Timothy R. Dreszer[1], Cricket A. Sloan[1], J. Seth Strattan[1], Forrest Y. Tanaka[1], Eurie L. Hong[4], and J. Michael Cherry[1,*]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA, and [2]Center for Biomolecular Science and Engineering, School of Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA

Project Website: https://www.encodeproject.org
Source Code: https://github.com/ENCODE-DCC/encoded, https://github.com/ENCODE-DCC/snovault
License: MIT

Main Text of Abstract

The Encyclopedia of DNA elements (ENCODE) project is an ongoing collaborative effort[1–6] to create a comprehensive catalog of functional elements initiated shortly after the completion of the Human Genome Project[7][1]. The current database exceeds 5500 experiments across more than 350 cell lines and tissues using a wide array of experimental techniques to study the chromatin structure, regulatory and transcriptional landscape of the *H. sapiens* and *M. musculus* genomes. All ENCODE experimental data, metadata, and associated computational analyses are submitted to the ENCODE Data Coordination Center (DCC) for validation, tracking, storage, and distribution to community resources and the scientific community. As the volume of data increases, the identification and organization of experimental details becomes increasingly intricate and demands careful curation. The ENCODE DCC[8–10] has created a general purpose software system, known as SnoVault, that supports metadata and file submission, a database used for metadata storage, web pages for displaying the metadata and a robust API for querying the metadata. The software is fully open-source, code and installation instructions can be found at: http://github.com/ENCODE-DCC/snovault/. The core database engine, SnoVault (which is completely independent of ENCODE, genomic data, or bioinformatic data) has been released as a separate Python package.