**Deciphering ENCODE**

Adam G. Diehl, Alan P. Boyle

**Introduction**

The primary goal of the ENCyclopedia of DNA Elements (ENCODE) project is at once simple and incredibly ambitious: to comprehensively annotate all functional sequences in the human genome. To add to this goal, ENCODE projects have also been launched focusing on the mouse, fly and worm genomes. To date, the consortium has released over 5,000 experiments, spanning nearly 300 cell and tissue types in human, mouse, fly and worm. In all, the repository houses more then 5 terabytes of data, around 20% of which is for mouse and a smaller, but growing, proportion is for fly and worm. With over 6,000 additional experiments currently proposed, the repository is projected to grow by nearly double by the end of 2016.

All this is possible due to the breadth of resources available to the ENCODE project. The key strength of the project stems from collaboration between numerous labs under the coordination of a panel of leading genomics experts from around the world. The result is unparalleled ability to rapidly produce, validate and deploy high quality genomics data. Participating labs utilize massively parallel sequencing and cutting edge computational technologies to streamline data production, processing and deployment. Assays are run according to standardized protocols spanning all stages of data production from cell growth through sequencing, and data are rigorously validated using a range of quality control metrics, ensuring that all data are of the highest quality attainable. With concerted efforts of the data coordination center (DCC), results are released to the public repository very quickly, and are freely and immediately available for use by anyone with internet access.

To date, ENCODE data have appeared in over 2,000 published scientific papers, more than half of which were conducted by investigators neither funded by nor associated with the ENCODE project, and interest continues to grow. However, the sheer size and complexity of the body ENCODE data can be intimidating, representing a possible entry barrier for prospective users. In this article, we hope to lower this barrier by providing practical background and examples of how ENCODE data is being used to augment hypothesis generation and validation in areas including disease genetics, pharmacogenomics, functional annotation, and comparative genomics. We will discuss what data are available, how and where to find them, and some caveats to keep in mind when selecting the most appropriate type(s) and production stage(s) of data for a given analysis.

We invite readers interested in finding out more about the background and rationale behind the ENCODE project to read the excellent discussions available in (ENCODE Project Consortium 2011; Pazin 2015), as well as the ENCODE summary papers for details of the individual projects’ results (ENCODE Project Consortium et al. 2007; ENCODE Project Consortium 2004; Harrow et al. 2012; Mouse ENCODE Consortium et al. 2012; Gerstein et al. 2010; modENCODE Consortium et al. 2010).

**ENCODE Data: What is available and where to find it**

The ENCODE project currently offers data spanning seven main categories (Figure??): 3D genome interactions, chromatin structure, DNA-protein interactions, DNA methylation, transcription, gene expression, and RNA-protein interactions. For each category, multiple labs following standardized protocols have contributed datasets from various high-throughput assays spanning a broad range of tissue and cell types. Individual datasets includes files from all production stages: raw sequencing reads, aligned sequence reads, finished data in one or more formats – typically including peak calls and/or genome-wide scores, and associated documentation. After validation by the DCC, datasets are entered into the ENCODE Repository, which organizes the data according to an extensive system of metadata. From there, datasets are available through the Web Portal and various affiliated outlets, each offering a different modality to navigate, view, and retrieve subsets of the data.

The first choice facing researchers interested in using ENCODE data is often which of these resources to use. This decision is largely based on two major factors: the genomic scope of the analysis, and the type(s) and amount of data needed. In terms of scope, we are concerned with whether the analysis targets a single, well-defined genomic interval or small set of intervals, a large set of intervals scattered throughout the genome, or focuses on genome-wide properties. In terms of the type(s) and quantity of data, this relates to the type of comparison – e.g., qualitative versus quantitative, and the method(s) applied. In particular, the type and production stage of data needed will affect the choice of source, as not all types and stages are available from all providers. Other factors to consider are whether a graphical view of the data is needed and the amount and type(s) of genomic context needed, as different outlets specialize in different type(s) of annotations. We will further discuss these choices in the following sections, including a practical discussion of three published studies, each using ENCODE data in different ways, in the Case Studies section, giving special attention to the choice of data for each.

**Choice of Access Point**

The ENCODE Web Portal (<https://www.encodeproject.org>) is the definitive source for ENCODE data and information regarding the samples, antibodies, standards and software used in their production. At present, raw data and files from intermediate processing stages are only available directly from ENCODE, thus, users requiring them must use either the Portal or API (discussed below). The Portal provides a powerful and flexible search tool and convenient methods to download whole datasets, making it well suited to genome-wide analyses. Search results include links to individual experiments, where their metadata and documentation can be found and individual files downloaded. Options are also available to download the entire set of search results in bulk or send them to the UCSC Genome Browser for visualization. This tool is particularly useful for analyses requiring genome-wide data for one or a small set of experiments, for example, comparing gene expression levels between human fetal and adult livers.

For users with more complex requirements, ENCODE offers a REST API (<https://www.encodeproject.org/help/rest-api/>) to query the repository programmatically. The API offers all the same search functionality as the Web Portal with the added power and flexibility of a programming language. This allows queries to be combined, subdivided, stored and manipulated in arbitrarily complex ways. As a result, the API is very useful for performing complex queries not possible using the Web Portal alone. For instance, queries seeking complex intersections spanning multiple cells, tissues, species and experiments are possible using the API but not the Portal. Although the API requires some programming expertise, it offers a simple and extensible way to perform complex queries and automate retrieval of large datasets.

For users focused on a defined genomic region (or set of regions) and who do not require access to the raw data, the UCSC and ENSEMBL genome browsers, HapMap, Roadmap Epigenomics and other sources (complete list at https://www.encodeproject.org/about/data-access/), offer various views of the processed ENCODE data embedded within their genomic context. Most offer both graphical browsers and extensible search platforms, such as the UCSC Table Browser. The graphical views offer intuitive ways to visually associate ENCODE data and other annotations within a single genomic region of interest. By contrast, the search platforms are useful for obtaining data associated with sets of genomic intervals, for instance, to test for enrichment of annotations within a set of target regions compared to randomly selected background regions. Many of these tools specialize in providing data of a specific type, and most are intuitive, easy to use and well-documented ((ENCODE Project Consortium 2011) (Rosenbloom et al. 2010)(<https://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html>) (hapmap, 1000 genomes, …) (<http://www.genome.gov/27553900>).

**Choice of Data**

There are many factors to consider in choosing the type(s) and production stage(s) of data for a given analysis. While it is beyond the scope of this article to comprehensively discuss all matters related to data selection, we offer some guidelines, particularly pertaining to the production stage from which to start.

When considering whether to use the fully-processed peak/score data or begin with the raw data or alignments, we are mainly concerned about how batch effects, non-biological effects that may quantitatively skew the results of massively-parallel sequencing experiments, might influence interpretation of the data. It has been well established that, even when experimental protocols are tightly controlled, these can be a significant source of quantitative variation between datasets produced under different conditions – e.g., in different labs or on different instruments (Li et al. 2014). This is particularly relevant to the ENCODE data, since so many labs have contributed to their production.

While quantitative comparisons of absolute signal levels, e.g., RNA-seq gene expression, are most likely to be affected by this phenomenon, peak-calling data may also be influenced by signal variations arising from batch effects. When using data from multiple experiments, users are advised to pay close attention to which lab(s) produced them, which protocol(s) and instruments were used in their production, and which specific programs, procedures and options were used to process the data. Many of these details can only be found in the documentation files accompanying each experiment, including specifics of the sample preparation and sequencing protocols, and specific programs and options used to prepare the final data. If there are differences between experiments used in an analysis, it is advisable to start with the raw or aligned sequencing data and apply a uniform processing pipeline, with parameters carefully chosen based on the production conditions indicated in the documentation for each experiment.

**Case Studies**

One common use of ENCODE data is to annotate and assign likely functions to SNPs affecting a disease or trait of interest. A recent study investigating blonde hair color in northern Europeans, provides an excellent example of how ENCODE data, coupled with carefully designed experimental follow-up, can be used to definitively locate a causal regulatory SNP (Guenther et al. 2014). Starting with a large GWAS dataset, the authors used transgenic methods to identify a segment of DNA exhibiting strong enhancer activity within a noncoding region ~350kb upstream of *Kitl,* a known pigmentation gene. This segment contained a previously annotated SNP and, by overlaying ENCODE transcription factor ChIP-seq data in the UCSC Browser, they were able to localize their SNP to a LEF1 binding site. By placing variants of this enhancer in a reporter construct, they showed that the SNP alters LEF1 binding affinity, causing a roughly 20% reduction in gene expression. Furthermore, placement of this sequence in an orthologous location in mice was sufficient to produce an observable pigmentation phenotype. Thus, the authors were able to identify the causal basis of a human phenotype and isolate it to a SNP within a previously uncharacterized enhancer lying over 350kb away from the target gene. The ENCODE data were able to suggest a testable hypothesis to extend the results of previous population genetics studies, leading to identification and validation of the causal SNP.

Our second example focuses on using ENCODE data to gain insights into genomic function. Embedded within the ENCODE data are signatures of broad regulatory principles and mechanisms. By integrating information from multiple ENCODE datasets, it is possible to find patterns that yield clues to such principles. One example is given by a recent investigation into how CTCF and the cohesin complex act to reproducibly form specific DNA loop structures, contributing to 3D genome organization (Guo et al. 2015). Using DNAse hypersensitivity, ChIA-PET, and CTCF ChIP-seq datasets obtained directly from ENCODE, they identified genomic regions that interacted through CTCF-tethered cohesin complexes, 78.7% of which included pairs of CTCF sites in an antiparallel arrangement. Through CRISPR-mediated inversion of a subset of sites, they were able to experimentally validate the importance of this organization, showing that inverting a single CTCF site can elicit specific, reproducible changes in loop topology. This is consistent with a loop-extrusion mechanism whereby opposing CTCF sites define loop boundaries (Nichols and Corces 2015). By combining the locations of open chromatin, CTCF binding, known 3D interactions, they were able to use the processed peak calls to elucidate a poorly understood mechanism underlying the basis of chromatin structures formation.

In our final example, Corradin and colleagues utilized multiple types of ENCODE data to investigate combinatorial GWAS SNP effects on gene expression (Corradin et al. 2013). Their first step was to identify putative enhancer elements, in 13 cell types, based on histone modification and DNAseI hypersensitivity data obtained from ENCODE and the Roadmap Epigenome Project. These were intersected with GWAS SNPs and correlated with various disease states in order to select a cell type and trait in which to analyze enhancer/SNP association patterns. The most significant associations were between six autoimmune disorders and enhancer-associated SNPs in GM12878 B-Lymphoblast-derived cells. When compared to non-disease-associated enhancer SNPs, enhancer SNP’s with autoimmune disease associations were significantly more likely to be in linkage disequilibrium (LD) with each other. This suggested a mechanism by which groups of SNPs, termed multiple-enhancer-variants (MEVs), act in combinatorion to elicit an effect. Consistent with this hypothesis, members of MEVs are more likely to target the same gene than expected by chance, and are significantly associated with differential expression between risk and non-risk individuals, while single-enhancer variants are not. Imperfect LD was found to decrease the strength of association with gene expression and, when individuals with imperfect LD were excluded, the linkage between the risk haplotype and gene expression became more significant, lending further support to the hypothesis that SNPs are acting cooperatively to affect gene expression. Tissue specificity of target genes was assayed using ENCODE RNA-Seq data from 11 human cell lines and NPC-cell-derived data from the Roadmap Epigenome Project. Results showed that genes targeted by GM12878-specific MEVs are significantly more likely to be tissue-specific and associated with immune-related GO terms than SEVs and non-disease-associated SNPs, consistent with roles in autoimmune disease. By combining multiple ENCODE datasets with data from the Roadmap Epigenome Project and HapMap), the authors were able to provide solid evidence for a mechanism by which noncoding SNPs contribute to disease susceptibility. In both the enhancer prediction and gene expression comparisons, they started with the raw sequencing data and applied careful normalization. As a result, they were able to incorporate datasets from disparate sources while minimizing the influence of batch effects.

**Example Application of the ENCODE API**

In our third case study, the authors performed a complex set of analyses using ENCODE and non-ENCODE data from several different cell types. In order to avoid bias, they needed to restrict their analyses at each step to cells for which comparable data were available in all lines. For example, in predicting enhancers, they needed to select tissues that had sufficient histone modification data and, within this set of tissues, select histone modifications for which data was available in all cell types. While conceptually simple, finding datasets that meet all these criteria using the Web Portal alone would require multiple steps and extensive manual curation. The REST API provides a way to automate these individual queries and find the datasets that overlap all groups using a simple algorithm. In this section, we will apply the API to a similar scenario: gathering the necessary data to compare global transcription factor ChIP-seq binding profiles between human K562 and mouse MEL cells.

Box 1 provides a pseudocode view of the program we have designed to find and retrieve this data. The first step is to perform discrete queries for transcription factor ChIP-Seq experiments in human and mouse. Search results are returned as JSON data, a tree-like format where metadata are stored as key:value pairs. This data can be easily parsed using off-the-shelf modules for many programing languages. In our second step, we use a simple algorithm to access the “target” attribute for each result, which corresponds to the transcription factor targeted in the ChIP-Seq experiment, and find all the targets that are present in both the human and mouse results. In the final step, we retrieve metadata for peak data files, in bigBed format, for each matching experiment, download the files and store the metadata for later identification. Using the power of the API, all of this can be accomplished with a single command, shown in Box 2. We are left with 118 files and their associated metadata, which we can use as the starting point for a comparative analysis of peak locations between the two species. While our example is relatively simple, it is easy to see how these methods can be extended to more complex analyses, such as those performed by Corradin and colleagues. We encourage readers to make use of the API and offer our scripts, through our github repository (URL), as a starting point. Box 3 contains a list of help resources and commonly used query parameters.

**The Future of ENCODE: Ensuring its Place at the Forefront of Genomics**

While the ENCODE project has proven its value in systematizing the production, storage and dissemination of genomics data, it has also been the subject of controversy. Recently, questions have been posed about the exact scope of its mission, its operational definition of functional sequence and whether it is appropriate for a “big science” project to posit on matters traditionally in the realm of “small science”(Eddy 2012; Doolittle 2013; Eddy 2013; Graur et al. 2013). Indeed, the ENCODE leadership has not always done the best job of embracing a role as a provider of primary data, although this purpose is arguably more defensible, if less exciting, than that of fundamentally reframing the definition of functional sequence. The temptation to extend the role of ENCODE increases, perhaps, as the cost of massively parallel sequencing decreases and it becomes more commonly available. Indeed, it is a matter for discussion whether a primary data provider role is sufficient to justify ENCODE’s existence. However, ENCODE is uniquely positioned in its ability to coordinate and centralize data production, processing and quality control efforts. The resources and expertise brought together by the consortium have already yielded many valuable insights into how best to produce and use high-throughput genomics data. With the continued support of the community, it will continue to do so, yielding further insights that will increase speed of production, quality and utility of the data: knowledge which will be of great use to the entire scientific community. Furthermore, because ENCODE data freely available to everyone, researchers are freed to focus on generating and testing hypotheses rather than on the details of data generation. These are just a few reasons the ENCODE Project maintains its relevance to the community and deserves of our continued support. However, as genomics technologies advance and our understanding of the scientific principles acting upon the genome evolve, so must ENCODE in order to stay relevant.

With its reliance on massively parallel sequencing, one of the biggest hurdles facing ENCODE relates to our improved understanding of batch effects: technical variables that quantitatively affect high throughput sequencing results. It is well accepted that, standardized protocols notwithstanding, batch effects can significantly influence the observed variation between samples, particularly those processed in different laboratories (Li et al. 2014). While it is possible to include various controls to explicitly separate biological and batch effects, much of the existing ENCODE data were generated before the importance of such measures was recognized and, even now, they are difficult to fully implement due to the distributed nature of the project. To our knowledge, there is currently no generally accepted method to normalize for batch effects in the absence of such control data. Attempts to work around these limitations (Lin et al. 2014; Gilad and Mizrahi-Man 2015) have been controversial and their adequacy remains a matter of intense debate. While the ENCODE leadership has given this considerable attention, it is not yet clear whether the measures under discussion by the consortium will be sufficient. Resolving this issue should remain a high priority as identifying the best way to deconvolute biological and batch effects in the absence of complex controls would both vastly improve the utility of existing ENCODE data and be of great value to the scientific community at large.

Another matter of great importance is improving off-the-shelf usability of the data. While ENCODE has done an admirable job of standardizing experimental protocols and sequencing, the data processing pipelines applied to the raw data are far more variable. In our experience, this can be a significant source of variation between datasets. Unfortunately, the time and expertise needed to process the raw data in-house represent a significant barrier, particularly to users lacking computational experience and/or access to the necessary computing resources. Furthermore, even for those with the experience and resources, the detailed information needed to choose appropriate parameters and options for programs used in the process can be hard to find. As noted earlier, critical details, if present at all, are often buried deep within the accompanying PDF documentation files, which appear neither to follow a standardized format nor explicit guidelines for what information must be included. In the short term, we suggest enforcing a uniform format for documentation and encoding the text directly as a field in the database to make these details visible to the API. Over the long-term, standardized analytical pipelines should be implemented and applied to new and existing data to obviate the need for external normalization.

Finally, while ENCODE’s stated goal is to be a comprehensive functional catalog, it has, naturally, been necessary to make decisions about which specific types of data, biological samples, developmental stages and methodologies to include. While the current state of the repository is indeed impressive, the consortium has just barely scratched the surface and the catalog is, at present, far from comprehensive. While filling in the gaps is an important objective, as new types of assays become available it becomes harder to prioritize where resources are best applied. It is critical for ENCODE to provide relevant data in a timely manner and so these goals must be carefully balanced against the realities of limited resources. In short, it is unlikely that ENCODE will ever realize its goal (admittedly, a moving target). However, as high-throughput genomics technology become more commonplace, increasing volumes of complementary genomics data are being produced in outside labs. While it is not, strictly speaking, within ENCODE’s operational mission, it would be a very attractive prospect to consolidate access to such outside datasets using the functionality of the ENCODE web portal and API. We believe that providing a way to integrate such datasets, provided they meet the same standards as the ENCODE data, would be highly valuable in bridging gaps in the ENCODE data and attracting interest in the repository as a public resource.

**Concluding Remarks**

Although we have provided just a glimpse of what is possible through use of the ENCODE data, we hope that we have been able to show its potential to enhance genomic analyses and aid in hypothesis generation and validation. While the body of ENCODE data is large and complex, the Consortium has invested heavily in developing resources to organize the data and make it approachable. From the Web Portal and API to the range of partners and third-party sites offering various views of the ENCODE data, tools are available for many specialized purposes. By carefully considering the scope and needs of an analysis and making use of a few simple principles, deciding on a data type, production stage and access point is reasonably straightforward. The data are neither perfect nor complete and there is room for improvement in many areas but none are insurmountable. On balance, the current limitations and controversies surrounding ENCODE are far outweighed by its merits. Continued efforts to address these challenges are ongoing and the repository is poised to expand exponentially, becoming even more useful as a resource for functionally annotating the genome. The number of publications using ENCODE data continues to grow, bringing many important insights on genomic functions, mechanisms and principles to the community. As the ENCODE resources and datasets improve, their ability to shed light on the many still unanswered questions will only increase.

**Bibliography**

Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal-Lari, R., Lupien, M., Markowitz, S. and Scacheri, P.C. 2013. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res*.

Doolittle, W.F. 2013. Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci U S A* 110(14), pp. 5294–5300.

Eddy, S.R. 2012. The C-value paradox, junk DNA and ENCODE. *Curr Biol* 22(21), pp. R898–R899.

Eddy, S.R. 2013. The ENCODE project: missteps overshadowing a success. *Curr Biol* 23(7), pp. R259–R261.

ENCODE Project Consortium 2011. A user’s guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* 9(4), p. e1001046.

ENCODE Project Consortium 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306(5696), pp. 636–640.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146), pp. 799–816.

Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., et al. 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* 330(6012), pp. 1775–1787.

Gilad, Y. and Mizrahi-Man, O. 2015. A reanalysis of mouse ENCODE comparative gene expression data. [version 1; referees: 3 approved, 1 approved with reservations]. *F1000Res* 4, p. 121.

Graur, D., Zheng, Y., Price, N., Azevedo, R.B., Zufall, R.A. and Elhaik, E. 2013. On the immortality of television sets: “function” in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* 5(3), pp. 578–590.

Guenther, C.A., Tasic, B., Luo, L., Bedell, M.A. and Kingsley, D.M. 2014. A molecular basis for classic blond hair color in Europeans. *Nat Genet* 46(7), pp. 748–752.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D., Jung, I., Wu, H., Zhai, Y., Tang, Y., Lu, Y., Wu, Y., Jia, Z., Li, W., Zhang, M., Ren, B., Krainer, A., Maniatis, T. and Wu, Q. 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* 162(4), pp. 900–910.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R. and Hubbard, T.J. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 22(9), pp. 1760–1774.

Li, S., Łabaj, P.P., Zumbo, P., Sykacek, P., Shi, W., Shi, L., Phan, J., Wu, P.Y., Wang, M., Wang, C., Thierry-Mieg, D., Thierry-Mieg, J., Kreil, D.P. and Mason, C.E. 2014. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat Biotechnol* 32(9), pp. 888–895.

Lin, S., Lin, Y., Nery, J.R., Urich, M.A., Breschi, A., Davis, C.A., Dobin, A., Zaleski, C., Beer, M.A., Chapman, W.C., Gingeras, T.R., Ecker, J.R. and Snyder, M.P. 2014. Comparison of the transcriptional landscapes between human and mouse tissues. *Proc Natl Acad Sci U S A* 111(48), pp. 17224–17229.

modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., et al. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* 330(6012), pp. 1787–1797.

Mouse ENCODE Consortium, Stamatoyannopoulos, J.A., Snyder, M., Hardison, R., Ren, B., et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* 13(8), p. 418.

Nichols, M.H. and Corces, V.G. 2015. A CTCF Code for 3D Genome Architecture. *Cell* 162(4), pp. 703–705.

Pazin, M.J. 2015. Using the ENCODE Resource for Functional Annotation of Genetic Variants. *Cold Spring Harb Protoc* 2015(6), pp. 522–536.

Rosenbloom, K.R., Dreszer, T.R., Pheasant, M., Barber, G.P., Meyer, L.R., Pohl, A., Raney, B.J., Wang, T., Hinrichs, A.S., Zweig, A.S., Fujita, P.A., Learned, K., Rhead, B., Smith, K.E., Kuhn, R.M., Karolchik, D., Haussler, D. and Kent, W.J. 2010. ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res* 38(Database issue), pp. D620–D625.

Box 1: Pseudocode for search\_human-mouse.pl

1. Process user inputs

* Check command line to see if we are downloading data files or only retrieving metadata
* Store specified file and output types, if supplied

1. Submit queries for human and mouse to ENCODE Portal and parse JSON response to find the range of experiments for each

* initialize hashes: %human\_factors and %mouse\_factors. Keys will be transcription factor symbols
* foreach species in (human, mouse)
  + build query URL
  + run query against ENCODE and store response as JSON object
  + loop over results in ‘@graph’ section of JSON
    - extract TF symbol, target, from result{target}
    - push a reference to the result in @{species{target}} array

1. Find the intersection between the human and mouse datasets

* initialize @intersection array
* for each factor in (keys(human))
  + if (exists(mouse{factor}))
    - push references to human{factor} and mouse{factor} to @intersection

1. Locate the files of interest and their metadata

* Initialize @metadata @files and @downloads arrays
* for each experiment in @intersection
  + get metadata for the current experiment
  + if metadata fulfills user-supplied constraints
    - if downloading data files
      * store files list in @files array
      * for each file in @files
        + get file metadata from ENCODE and store as JSON
        + check JSON data for any/all of: (output\_type, file\_format, file\_format\_type); if these match user-supplied criteria:

store metadata in @metadata

store file JSON in @downloads

* + - * + else move on to the next file
    - else store metadata in @metadata
  + else move on to the next record

1. Download matching experiments and/or print formatted metadata

* if downloading files
  + for each file in @downloads
    - get file download URL from file JSON
    - download the file
* for each row in @metadata, print a row to the metadata file

Box 2: search\_human-mouse.pl command and output

Command to retrieve peak locations for all transcription factors with ChIP-Seq data in human and mouse:

./search\_human-mouse.pl K562 MEL "&assay\_term\_name=ChIP-seq&target.investigated\_as=transcription factor" --out-root chipseq --download --output-type peaks --file-format bigBed

Arguments: Human Cell (K562), Mouse Cell (MEL), “query parameters string” – see Box 3 for definitions

Options Used:

--download

Download files associated with the results. (Only metadata is saved without this)

--output-type peaks

Limits results to files of given type(s). File types are

matched against the "output\_type" column of the file records, and

available values will vary depending on the experiment.

--file-format bigBed

Restrict downloads to a given file format.

--out-root <root>

String to prepend to output file names.

Truncated Program Output:

Query URL: http://www.encodeproject.org/search/?searchTerm=K562&replicates.library.biosample.donor.organism.scientific\_name=Homo sapiens&type=experiment&assay\_term\_name=ChIP-seq&target.investigated\_as=transcription factor&limit=all&frame=object&format=json

Success: 215 results found for Homo sapiens.

Query URL: http://www.encodeproject.org/search/?searchTerm=MEL&replicates.library.biosample.donor.organism.scientific\_name=Mus musculus&type=experiment&assay\_term\_name=ChIP-seq&target.investigated\_as=transcription factor&limit=all&frame=object&format=json

Success: 50 results found for Mus musculus.

Finding intersecting terms...

Found 32 terms with data in both species.

Retrieving data for 104 experiments...

Found a matching record at https://www.encodeproject.org/files/ENCFF000YGD/@@download/ENCFF000YGD.bigBed. Retrieving data...

Verifying Checksum...

Saving file to chipseq.ENCFF000YGD.bigBed...

…

Retrieved 118 files for 104 experiments.

Metadata written to chipseq.metadata

Done

search\_human-mouse.pl and a more generalized script, search\_encode.pl, are available for download at our github repository (url). Both are self-documented through the --help option and are freely available to use and modify under the terms of the GNU GPL.

Box 3: API Resources

Commonly Used Search Parameters:

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Description** | **Common Values/Format** |
| assay\_term\_name | type of assay | ChIP-seq  RNA-seq  DNase-seq  ChIA-PET  FAIRE-seq |
| assembly | Genome assembly referenced | hg19, mm9 |
| target | Target of ChIP-Seq assay | Any histone modification, transcription factor, etc. |
| target.investigated\_as | Type of ChIP-Seq assay | “transcription factors”  “histone”  “histone modification”  “RNA binding protein”  “control” |
| replicates.library.nucleic\_acid\_term\_name | Type of library | RNA  “polyadenylated mRNA” |
| replicates.library.biosample.  biosample\_type | Type of sample | “immortalized cell line”  tissue  “primary cell”  “in vitro differentiated cells”  “stem cell” |
| replicates.library.biosample.donor.  organism.scientific\_name | Scientific name of target species | “Homo sapiens”  “Mus musculus” |
| replicates.library.biosample.donor.  life\_stage | Developmental stage | Adult  child  fetal  embryonic  postnatal |
| searchTerm | Query term | Free-form text |
| limit | Number of results to show per page | all: show all results  *N*: show *N* results |
| type | Type of record | experiment  assay  biosample  antibody |
| lab.title | Laboratory in which data was produced | “Firstname Lastname, Institution” |
| files.file\_type | Show experiments for which this type of file is available | fastq, bam, wig, bigWig, gtf, bed, bigBed, tsv |
| files.run\_type | Type of sequencing run | single-ended  paired-ended |

API-Specific Parameters

|  |  |
| --- | --- |
| **Parameter** | **Description** |
| &format=json | Return search results as a JSON object |
| &frame=object | Include all database attributes in the results |

Tip: These parameters can be added/removed to/from the URL within the Web Portal to see how they affect the search results (in form of “&parameter=value”). Parameter names and values are case sensitive.

ENCODE API Help Section: https://www.encodeproject.org/help/rest-api/

ENCODE Database Schema: <https://www.encodeproject.org/profiles/graph.svg> -- Describes the fields and relationships between tables in the repository database. In many cases, non-standard search parameters can be built from these using the ‘.’ delimited format seen in the common parameters table.