**Deciphering ENCODE**

Adam G. Diehl, Alan P. Boyle

**Introduction**

The primary goal of the 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. With numerous participating labs within and outside the United States and leveraging of massively parallel sequencing and cutting edge computational technologies, data production, processing and deployment are rapid. Participating labs follow standardized protocols for cell and tissue growth, sample preparation and sequencing, and data are rigorously validated using a range of quality control metrics, ensuring that all data, regardless of source, are of the highest quality attainable. With concerted efforts to validate and catalog data by the data coordination center (DCC), results are released to the public 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 can being used to augment hypothesis generation and validation in areas such disease genetics, pharmacogenomics, functional annotation and comparative genomics. In the process, 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 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 data for various types of high-throughput assays spanning a broad range of tissue and cell types. Raw sequencing data are mapped and processed into genome-wide peak calls and/or scores and the complete dataset, including raw data, mapped data alignments, finished data in one or more formats, and associated documentation, is then submitted to the DCC for validation and distribution through the ENCODE Web Portal and various affiliated sources. Each outlet offers a different modality to navigate, view and retrieve subsets of the data and the first choice facing researchers interested in using the ENCODE data is often which of these resources to use. This decision often boils down to two major factors: the genomic scope of the analysis – i.e., whether it targets a single, well-defined genomic interval, a large set of intervals throughout the genome, or focuses on genome-wide properties; and the type(s) and amount of context necessary to view in conjunction with the ENCODE data. Other factors to consider are what type of data is needed, e.g., position-wise scores versus peak locations, and what production stage represents the best starting point, i.e., finished data versus raw or mapped reads. We will discuss each of these choices in the following two sections. In the Case Studies section, we will synthesize this information through discussion of three published studies, each employing the ENCODE data in different ways, 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. Through its faceted search functionality, the Portal provides the ability to browse ENCODE data based on its associated metadata, which is organized in a relational database. Once located, data may be downloaded manually by visiting the search result links for individual experiments, in bulk by clicking the “Download” button, or sent to the UCSC Genome Browser for visualization. This tool is particularly useful when genome-wide data for one or a small set of experiments are needed for an analysis, for example, to compare gene expression levels between using RNA-seq data from human fetal and adult livers. The Portal provides a powerful and flexible search tool and convenient methods to download whole datasets, making it well suited to genome-wide analyses. However, aside from integration with the UCSC browser, it does not offer a graphical view of the data nor much by way of genomic context.

In addition to the Web Portal, ENCODE offers a REST API (<https://www.encodeproject.org/help/rest-api/>) to interact with the repository programmatically. The API offers all the search functionality available through the Web Portal with the added power and flexibility of being able to combine and subdivide queries by storing and manipulating them using any programming language capable of sending and receiving HTTP data. The API can be very useful for locating and retrieving broad and complex datasets that span multiple experiments, and for automating retrieval of large amounts of data. There is a learning curve involved in using the API, particularly for those lacking experience with computer programming. However, it offers a simple and extensible way to perform complex queries against the datasets and can be the best way to find datasets for large comparative projects and others requiring data spanning a broad range of conditions and/or sources.

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, which offer an intuitive way to visually associate ENCODE data with other genomic features of interest, and extensible search platforms, the UCSC Table Browser for example, which are very useful for retrieving ENCODE data and other annotations spanning a set of genomic intervals, for instance, to test for enrichment of annotations within a set of genomic intervals compared to randomly selected background regions. Many of these tools specialize in providing data of a specific type, HapMap for population genetics data, for instance. These resources are, by and large, intuitive, easy to use and well-documented, with excellent descriptions and tutorials available ((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 or score data or to 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, may 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 in different labs, on different instruments, etc. (Li et al. 2014). This is particularly relevant to the ENCODE data, given its distributed nature.

While quantitative comparisons of the absolute level of signal, e.g., RNA-seq expression levels, are most likely to be affected, peak calling algorithms may also be influenced by signal variations arising from batch effects. Whenever utilizing data from multiple experiments, users are advised to pay attention to which lab(s) they originate from and, even within a lab, which protocol(s) were used in their production and processing, consulting the documentation files as necessary. These files contain low-level details of the sample preparation and sequencing protocols, and information on programs and options used to prepare the final data. If there are differences, it is advisable to start with the raw or aligned sequencing data and apply a uniform normalization and quantification 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 that SNP to a LEF1 binding site. By placing variants of this enhancer in a reporter construct, they were able to show that the SNP alters LEF1 binding affinity, causing a roughly 20% change in gene expression level. Furthermore, placement of this sequence in an orthologous location in mice was enough 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. In this case, 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 the signatures of broad regulatory principles and mechanisms. By integrating information from multiple ENCODE datasets, it is possible to find patterns that give us clues to these 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 the 3D organization of the genome (Guo et al. 2015). Using DNAse hypersensitivity, ChIA-PET and CTCF ChIP-seq datasets obtained directly from the ENCODE Portal, they identified genomic regions that interacted through CTCF-tethered cohesin complexes, finding that 78.7% of these included pairs of CTCF sites situated in an antiparallel orientation. Through CRISPR-mediated inversion of some of these sites, they were able to experimentally validate the importance of this organization, showing that inversion of a single CTCF site was sufficient to elicit observable changes in loop topology. This is suggestive of 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 make use of the processed peaks to elucidate a poorly understood mechanism underlying the basis of chromatin structures formation.

In our final example, Corradin et al (Corradin et al. 2013) utilized multiple levels of ENCODE data in their investigation of combinatorial GWAS SNP effects on gene expression. Raw ENCODE and Roadmap Epigenome Project histone modification and DNAse I hypersensitivity data were obtained, mapped, and normalized prior to making peak calls used to identify putative enhancer elements in 13 cell types. These were intersected with GWAS SNPs and used to determine correlations with various diseases in order to select a cell type and trait suitable for their analysis of enhancer/SNP association. They observed a significant correlation between six autoimmune disorders and SNP-associated enhancer regions in GM12878 B-Lymphoblast-derived cells. Further analysis showed that, among enhancer-associated SNPs, those associated with autoimmune disease were more frequently in linkage disequilibrium (LD) with each other relative to non-disease-associated SNPs. This suggested a mechanism by which groups of SNPs, termed multiple-enhancer-variants (MEVs), act in combination to produce an effect. Consistent with this hypothesis, these groups of linked enhancers are more likely to target the same gene than expected by chance. Based on HapMap B-lymphocyte gene expression data, they demonstrated that MEVs are significantly associated with differential expression between risk and non-risk individuals, while single-enhancer variants are not. Furthermore, when individuals with imperfect LD are excluded, this association becomes even more significant, lending support to their hypothesis. To tie these expression effects more closely to the relevant disease phenotypes, ENCODE RNA-Seq data from 11 human cell lines and NPC-cell-derived data from the Roadmap Epigenome Project were used to investigate their tissue-specificity. Starting with the raw sequencing reads, they quantified and normalized the expression values observed in each tissue type. Results showed that genes targeted by GM12878-specific MEVs are indeed significantly more likely to be GM12878-specific in expression, and are enriched for immune-related functions based on their associated GO terms. By combining multiple types of ENCODE data with data from the Roadmap Epigenome Project and HapMap, through careful mapping, quantification and normalization of the raw datasets, the authors were able to provide solid evidence for one mechanism by which noncoding SNPs contribute to disease susceptibility.

**Example Application of the ENCODE API**

The ENCODE Web Portal is a powerful and flexible tool for exploring the ENCODE data but it lacks certain capabilities, including the ability to compare multiple searches (to find the intersection, for instance), and the ability to download only a single file type for a set of search results. The REST API answers these limitations by incorporating the power and flexibility of a programming language, allowing queries to be combined, subdivided and refined to generate complex “superqueries”. Results are returned as JSON data, a tree-like format where key:value pairs are stored in a nested hierarchy, which can be parsed using readily available modules for Perl, Python and many other popular programming languages, and can be store and manipulated in numerous ways. Because it interfaces with the repository database through the same HTTP framework as the Web Portal, API queries can be conveniently (p)reviewed by plugging the search URL directly into the Web Portal. In this section, we will work through an example of how to use the API to find the intersection between two searches: something that is not possible through a simple Web Portal query.

Our task is to gather 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 perform this search, search\_human-mouse.pl. Briefly, our strategy is first, to perform discrete queries for transcription factor ChIP-Seq experiments in human and mouse, then second, find the intersection between factors in those result sets, and finally, retrieve bigBed peak files for each matching experiment. Using the power of the API, all of this can be accomplished with a single command, shown in Box 2, which retrieves a set of 118 data files, and the file “chipseq.metadata”, containing a row of metadata that identifies each file and relates it back to its parent experiment. With these data in-hand, many diverse analyses are possible and we will leave these to the reader’s imagination.

The above example clearly illustrates the power of the API to simplify complex data retrieval tasks. Gathering the same dataset through individual Web Portal queries would have required time-consuming and error-prone manual comparison of multiple search results. By contrast, using the API, we are able to accurately and reproducibly gather the necessary data with just a few keystrokes. We encourage readers to make use of the API, offering our scripts through our github repository (URL) as a starting point for experimentation, and a list of help resources and commonly used query parameters in Box 3.

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

While the ENCODE project has more than proven its value in systematizing the production, storage and dissemination of genomics data, it has also been the subject of controversy. Recently, some have posed questions 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 is increased, perhaps, as the cost of massively parallel sequencing decreases and its availability increases. 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 and will continue to yield valuable insights into how best to produce and use high-throughput genomics data. The ability of ENCODE to rapidly produce and deploy data to the public is still unparalleled and its existence democratizes the scientific process, facilitates innovation and frees researchers to focus on generating and testing hypotheses. These advantages should be sufficient to secure its place in the universe of genomics. That said, as technology and our understanding of the scientific principles acting upon the genome evolve, so must ENCODE in order to meet the requirements of scientists now and into the future.

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 that can properly 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 that have been planned/implemented will be sufficient. We hope this will remain an area of active investigation within the consortium as identifying the best way to deconvolute biological and batch effects in both future and existing datasets is not only crucial to the continued viability of ENCODE, but would on its own be a valuable contribution 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. 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 such that all are on even footing.

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 only scratched the surface of what is possible through clever 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 keep the data organized and make it approachable. From the Web Portal and API to the many partners and third-party sites offering various views of the ENCODE data, there is a tool available for every specialized purpose. 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. While the data are neither perfect nor complete and there is room for improvement in many areas -- particularly in addressing the influence of batch effects, standardizing analytical pipelines and documentation, and ensuring completeness of experimental metadata, none of these issues are insurmountable and, on balance, they are far outweighed by the merits of the ENCODE resources. As efforts to address these challenges continue and the repository expands, the utility of the ENCODE repository as a resource for functionally annotating the genome and investigating the principles by which it functions and evolves will only grow. The value of the ENCODE data to researchers in a broad range of disciplines is further proven by the growing number of non-ENCODE publications making use of the data and we are confident that this trend will continue.

**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 the user inputs

Check command line options 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 response JSON data 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

loop over results in ‘@graph’

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

foreach 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 and @downloads array

foreach 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

foreach file in @files

get json data for file from ENCODE

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

foreach file in @downloads

get file location from file json

download the file

foreach row in @metadata, print 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.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. Advanced users can leverage the relationships described to structure arbitrary search parameters using the ‘.’ delimited format shown in the common parameters table.