Trends in Genetics

CellPress

Review

Deciphering ENCODE

Adam G. Diehl[1] and Alan P. Boyle[1,2,*]

The ENCODE project represents a major leap from merely describing and comparing genomic sequences to surveying them for direct indicators of function. The astounding quantity of data produced by the ENCODE consortium can serve as a map to locate specific landmarks, guide hypothesis generation, and lead us to principles and mechanisms underlying genome biology. Despite its broad appeal, the size and complexity of the repository can be intimidating to prospective users. We present here some background about the ENCODE data, survey the resources available for accessing them, and describe a few simple principles to help prospective users choose the data type(s) that best suit their needs, where to get them, and how to use them to their best advantage.

Scope and Mission of the ENCODE Project

The primary goal of the Encyclopedia of DNA Elements (ENCODE) project is both simple and incredibly ambitious: to comprehensively annotate all functional sequences in the human genome. To add to this goal, ENCODE projects have been launched focusing on the mouse, fly, and worm genomes. To date the consortium has released over 5000 experiments, spanning nearly 300 cell and tissue types in human, mouse, fly, and worm. In all, the repository houses more than 5 terabytes of data, of which around 20% is for mouse and a smaller, but growing, proportion is for fly and worm. With over 6000 additional experiments proposed, the repository is projected to nearly double in size by the end of 2016.

All this is enabled by 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. The result is an 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 (Table S1 in the supplemental information online). With concerted efforts of the data coordination center (DCC), results are released to the public very quickly, and are freely and immediately available to anyone with internet access.

To date, ENCODE data have appeared in over 2000 peer-reviewed papers, more than half of which were authored by investigators outside the ENCODE project, and interest continues to grow. However, the sheer size and complexity of the 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 information and examples of how ENCODE data are being used to augment hypothesis generation and validation in 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.

Trends

High-throughput sequencing data have become the standard in providing evidence of function for genomic features of interest, offering a powerful way to develop hypotheses as to how they contribute to phenotypes.

The ENCODE repository offers data pertinent to a broad range of genomic analyses, facilitating insights into specific phenotypes as well as genome-wide properties and principles.

Despite their growing popularity, the size and complexity of the ENCODE repository can make it difficult to choose the best resources to support a given inquiry. However, considering the genomic scope and nature of an analysis can considerably simplify the choice of appropriate data.

ENCODE provides a range of tools, including a web search portal, API, and integration with several popular genome browsers, to help researchers navigate the resources and retrieve the necessary data for an analysis.

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
[2]Department of Human Genetics, University of Michigan, Ann Arbor MI 48109, USA

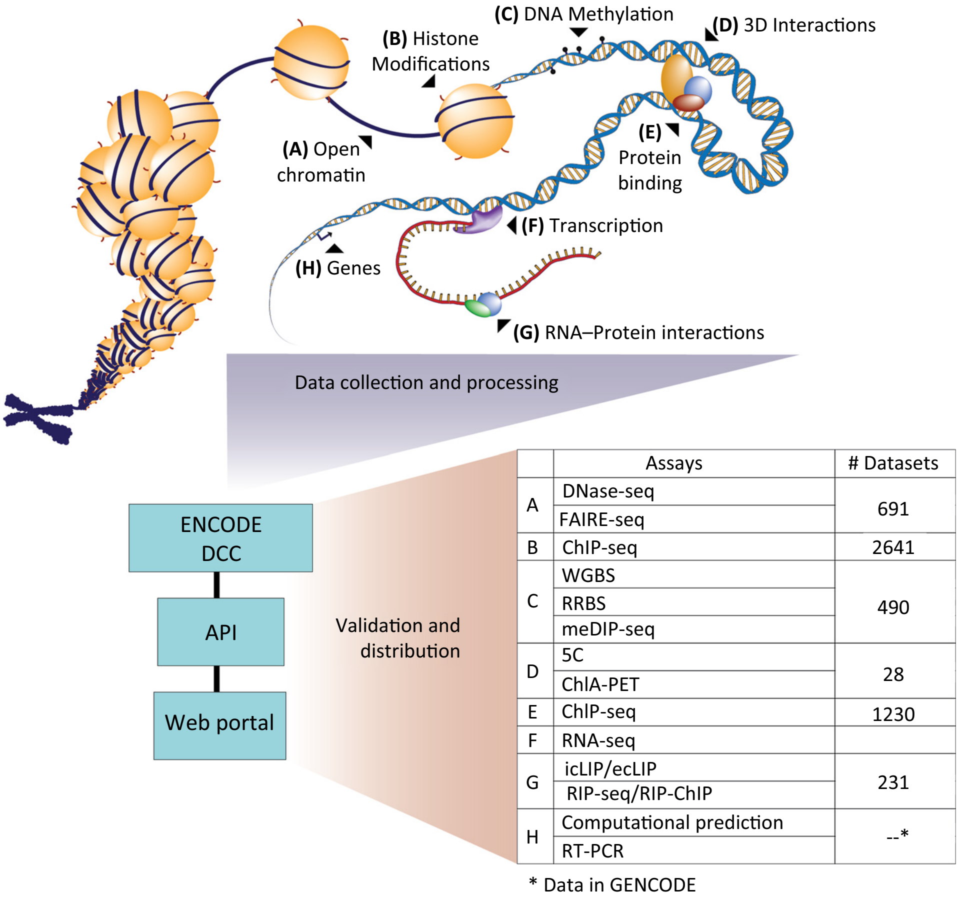
*Correspondence: apboyle@umich.edu (A.P. Boyle).

http://dx.doi.org/10.1016/j.tig.2016.02.002

We invite readers interested in learning more about the background and rationale behind the ENCODE project to read the excellent discussions in [1] and [2], and the ENCODE summary papers for information regarding individual projects [3–8].

ENCODE Data: What Is Available and Where To Find It

The ENCODE project currently offers data spanning seven primary categories (Figure 1): 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 include 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



	Assays	# Datasets
A	DNase-seq	691
	FAIRE-seq	
B	ChIP-seq	2641
C	WGBS	490
	RRBS	
	meDIP-seq	
D	5C	28
	ChIA-PET	
E	ChIP-seq	1230
F	RNA-seq	
G	icLIP/ecLIP	231
	RIP-seq/RIP-ChIP	
H	Computational prediction	--*
	RT-PCR	

* Data in GENCODE

Trends in Genetics

Figure 1. An Illustrative Example of the Data Types Available through the Encyclopedia of DNA Elements (ENCODE) Project Portals. These include measurements of (A) Open chromatin using DNase-seq and FAIRE-seq, (B) ChIP-seq for histone modifications, (C) DNA methylation, (D) 3D interactions, (E) ChIP-seq for transcription factors and other chromatin-associated proteins, (F) transcriptional output including RNA-seq, CAGE, and RNA-PET, (G) RNA–protein interactions, and (H) gene body predictions through computational and manual annotation as part of GENCODE. Abbreviations: 5C, chromosome conformation capture carbon copy; CAGE, cap analysis of gene expression; ChIA, chromatin interaction analysis; ChIP, chromatin immunoprecipitation; cLIP; cross-linking immunoprecipitation; FAIRE, formaldehyde assisted isolation of regulatory elements; meDIP, methylated DNA immunoprecipitation; PET, paired-end tag; RIP, RNA immunoprecipitation; RRBS, reduced representation bisulfite sequencing; seq, sequencing; WGBS, whole-genome bisulfite sequencing.

documentation. After validation by the DCC, datasets are entered into the ENCODE repository, which organizes the data according to an extensive system of metadata.

Datasets in the repository are available through the Web Portal and several affiliated outlets, each offering a different modality to navigate, view, and retrieve ENCODE data. The first choice facing researchers interested in using the data is which of these resources to use, a decision largely based on two factors: the genomic scope of the analysis and the type(s) of data needed. The genomic scope relates to whether the analysis focuses on the individual properties of a small number of defined genomic intervals, a larger set of intervals scattered throughout the genome, or focuses on genome-wide properties inferred from numerous regions, potentially employing multiple annotations. The ideal type and production stage of data needed is dictated by the nature of the inquiry. For example, a study comparing absolute levels of gene expression may require raw sequencing data as its input (for reasons discussed in the Choice of Data section), and these are only available from the ENCODE portal. Other considerations are whether a graphical view of the data is needed, if it is more advantageous to directly access the ENCODE data or rely on a specialized resource offering functional annotations based on multiple ENCODE datasets (e.g., RegulomeDB [9]), and if additional annotations from outside ENCODE are necessary. In the following sections we present specific considerations for choosing the right data, including practical examples and a discussion of three case studies, each using ENCODE data in different ways, giving special attention to the choice of data used in each.

Choice of Access Point

The ENCODE web portal (www.encodeproject.org) is the definitive source for ENCODE data and information regarding the samples, antibodies, standards, and software used in their production. The portal provides a powerful search tool and convenient methods to download datasets in bulk, making it well suited to genome-wide analyses. Search results include links to individual experiments where their metadata, documentation, and data files can be found and downloaded. Options are also available to download entire sets of data files from a query or to send the results to the University of California Santa Cruz (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. Currently, raw sequencing data and files from intermediate processing stages can only be accessed through the portal or application program interface (API).

For users with more complex requirements, ENCODE offers an API (www.encodeproject.org/help/rest-api/) to query the repository programmatically. The API offers the same 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 to achieve specific search goals. Thus, the API is very useful for performing complex queries not practical to perform using the web portal alone. For instance, queries seeking intersections between multiple targets, cells, tissues, and conditions are greatly simplified using the API. Although it requires some programming expertise, the API offers a simple and extensible way to perform complex queries and automate data retrieval.

When a dataset consists of a small set of defined genomic regions, the graphical browsers offered by UCSC ENSEMBL, HapMap, Roadmap Epigenomics, and several other sources (complete list at www.encodeproject.org/about/data-access/) may be ideal. These offer the ENCODE datasets in the context of genomic sequence and diverse outside annotations, making them very powerful for screening discrete genomic regions, such as single-nucleotide polymorphism (SNP) locations, for functional evidence. For larger sets of intervals, search platforms (e.g., the UCSC table browser) offered by these resources may be ideal. These offer the functionality to intersect sets of regions with multiple annotation tracks and download the results for local analysis. For example, a set of target and background regions could be intersected with histone modification data to test for enrichment of putative enhancers in a dataset. These tools are well documented in the primary literature (e.g., [1,10–13]), and most also provide extensive online help resources.

An alternative for users with modestly sized datasets is to utilize a tool that summarizes multiple ENCODE datasets to provide specialized functional annotations. For instance, RegulomeDB combines ENCODE histone modification, DNase hypersensitivity, and transcription factor ChIP-chip data to rank and classify noncoding SNPs according to their likelihood of affecting regulatory function [9]. Similar tools for a variety of specialized purposes are available, and can be found by searching the 'software tools' subsection of the ENCODE publications list.

Choice of Data

There are many factors to consider in choosing the type(s) and production stage(s) of data for an 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 data, ChIP-seq peak locations, for example, or to begin with an earlier production stage, such as the aligned sequence reads, the main concerns is how batch effects (non-biological effects that may quantitatively skew the results of massively-parallel sequencing experiments) might influence the interpretation of the data. It has been well established that, even when experimental protocols are tightly controlled, these can be a significant source of variation between datasets produced under different conditions, such as in different labs or on different instruments [14]. This is particularly relevant to the ENCODE data because so many labs have contributed to their production.

Because batch effects influence absolute signal levels, comparisons relying directly on these measurements (e.g., RNA-seq gene expression) are most prone to batch effects. However, because these signal variations can affect the locations and boundaries of peak predictions, batch effects may also influence location-based comparisons, especially if they are sensitive to differences in individual peak boundaries. Thus, when making use of multiple datasets, all users should consider the lab of origin, library preparation and sequencing protocols, and the instruments used in producing the datasets. If any of these differ substantially, significant batch effects may exist even if all data originate from the same lab. One option is to select only datasets close enough in provenance to be minimally affected, but if this is not possible it is advisable to apply some type of normalization. Depending on the nature and stringency of the analysis, this may be as simple as quantile normalization of the final scores or as elaborate as the multistep process used in a recent reanalysis of tissue-based gene expression in human and mouse [15]. Some options and considerations are discussed in [14,16–24].

Even when data were produced in the same lab, they may have been processed using different analysis pipelines to produce the final data, which can also be a source of significant variation. In these cases, one option is to reprocess the data using the ENCODE uniform data processing pipelines (www.encodeproject.org/pipelines/). We provide a list of common tools, their descriptions, and references used in the ENCODE pipelines in Table S1. Finally, it is important to note that many of the details mentioned in this section are often absent from the metadata of an experiment, and it is therefore crucial to consult the documentation files for each experiment when making these considerations.

Case Studies

ENCODE data have been used in many studies for numerous purposes: to guide hypothesis formation, infer biological properties of the genome, support the functional significance of a genomic region(s), or infer the functions of unclassified genomic features, for example. Although it is not practical to give a comprehensive review of all >2000 published uses of ENCODE data, we will explore three case studies to illustrate how different types of analyses affect the choices of data (i.e., source, type, and production stage) and how they are handled (i.e., off-the-shelf versus processed in-house). We provide the reader with examples of additional studies using ENCODE data for diverse analyses in Table S2.

Case 1 Causal SNP Identification

ENCODE data are frequently used to annotate and assign likely functions to SNPs affecting a trait of interest. A recent study investigating blonde hair color in northern Europeans, for instance, demonstrates how ENCODE data, coupled with experimental follow-up, can be used to definitively locate causal regulatory SNPs [25]. Starting with a large genome-wide association study (GWAS) dataset, the authors used transgenic methods to identify a DNA sequence exhibiting strong enhancer activity within a noncoding region ~350 kb upstream of *Kitl*, a known pigmentation gene. This sequence contains a known SNP and, by overlaying ENCODE transcription factor ChIP-seq data in the UCSC browser, they identified an overlapping LEF1 transcription factor 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 350 kb away from the target gene. The ENCODE data associations suggested a testable hypothesis to extend previous population genetics studies, leading to identification and validation of the causal SNP. An alternative approach would have been to use a resource such as RegulomeDB [9] or haploReg [26], both of which include the association between this SNP and TCF/LEF binding, although neither annotates it as having strong evidence of function.

Case 2 Investigating Global Genomic Mechanisms

Our second example focuses on mining ENCODE data for the embedded signatures of global regulatory principles. By integrating information from multiple ENCODE datasets it is possible to find clues to such principles, as illustrated by a recent investigation of how the mammalian CCCTC-binding factor (CTCF) and the cohesin complex contribute to 3D genome organization [27]. Previous work suggested a role for convergent, antiparallel arrangement of CTCF sites in defining chromatin loop boundaries. Through CRISPR-mediated inversion of one site in an interacting pair, the authors showed that perturbing this pattern disrupts chromatin looping between enhancers and promoters (defined using ENCODE DNase hypersensitivity and histone-modification data), leading to decreased target gene expression, in two well-studied mammalian gene clusters. By integrating ENCODE CTCF ChIP-seq and ChIA-PET data, they were able to extend these findings to the whole genome. CTCF peaks were intersected with the ChIA-PET data to identify sites that interact as part of CTCF-tethered cohesin structures, of which 78.7% shared this convergent, antiparallel arrangement. These comprise >90% of CTCF/cohesin-mediated chromatin domains and are present at >60% of topologically associated domain boundaries, suggesting the importance of this arrangement in directing global 3D structure. In a companion paper, Nichols and Corces elaborated on their findings, suggesting a loop-extrusion mechanism whereby opposing CTCF sites define chromatin loop boundaries [28], a tantalizing hypothesis to direct future studies into the mechanisms underlying 3D genome organization.

Case 3 Investigating Mechanisms of Gene Regulation Using ENCODE and Non-ENCODE Data

In our final example, Corradin and colleagues combined multiple types of ENCODE data with complementary datasets from other sources to investigate combinatorial GWAS SNP effects on gene expression [29]. This analysis was performed in three stages, and a key determinant in its success was the application of careful selection and normalization of raw data at each stage to minimize the influence of batch effects. The first stage involved selecting an appropriate cell line for further analysis. Putative enhancer elements were identified in 13 cell types based on histone modification and DNase hypersensitivity data from the ENCODE and Roadmap Epigenome Projects. These predictions were intersected with GWAS SNPs and correlated with various disease states, and the cell line with the most significant associations, GM12878 B lymphoblast cells, was selected. In the second stage, the relationships between individual enhancer SNPs (enhSNPs), and between enhSNPs and their target genes, were compared within this cell line. The observed enrichment of autoimmune disease-related enhSNPs in linkage disequilibrium (LD) with each other compared to non-disease-associated (neutral) enhSNPs suggested a mechanism whereby SNPs within multiple, distinct enhancers (MEVs) act in combination to elicit an effect. Consistent with this hypothesis, members of MEVs are more likely to target the same gene than single-enhancer variants (SEVs), and are significantly associated with differential expression between risk and non-risk individuals, whereas SEVs are not. Furthermore, excluding individuals with imperfect LD in MEVs increased the strength and significance of the association with differential gene expression. In the final stage, tissue specificity was used to assess the relevance of genes targeted by MEVs to autoimmune disease. ENCODE RNA-Seq data from 11 human cell lines and neural precursor cell (NPC)-derived data from the Roadmap Epigenome Project were normalized and correlated with the three classes of enhSNPs. Results showed that genes targeted by GM12878-specific MEVs are significantly more likely to be immune-specific and associated with immune-related Gene Ontology (GO) terms than SEVs or neutral SNPs, consistent with roles in autoimmune disease. Taken together, these findings strongly support a mechanism whereby disease susceptibility depends on genetic variation at multiple enhancers; by combining ENCODE datasets with complementary data from other sources the authors were able to provide solid evidence for this hypothesis.

Example Application of the ENCODE API

In Case 3, the authors performed a complex set of analyses using multiple lines of ENCODE and non-ENCODE data. At each step, it was crucial that they restrict their analyses to cells for which comparable data were available in all lines. For example, in predicting enhancers, they needed to select cells with sufficient histone-modification data and, within this set, select histone modifications for which data were available in all cell types. While conceptually simple, finding datasets that meet these criteria using the web portal alone would require multiple steps and extensive manual curation. The API offers a way to automate queries to find datasets satisfying complex criteria. In this section we will apply the API to a similar problem: gathering the necessary data to compare global transcription factor ChIP-seq binding profiles between human K562 and mouse MEL cells.

Box 1 provides pseudocode for a program designed to find and retrieve these data. The first step is to perform discrete queries for transcription factor ChIP-Seq experiments in human and mouse. Search results from ENCODE are returned as JSON-formatted metadata, which represent various attributes of the datasets as key:value pairs. These can be easily parsed using standard modules for many programming languages, as in the second step, where a simple algorithm retrieves the 'target' attribute, corresponding to the transcription factor targeted in the ChIP-Seq experiment, for each result, and finds the intersection between the target lists for human and mouse. Finally, we retrieve metadata for individual experiments, from which URLs for the bigBed peak files are found. For each record found, the file is downloaded

Box 1. Pseudocode for search_human-mouse.pl

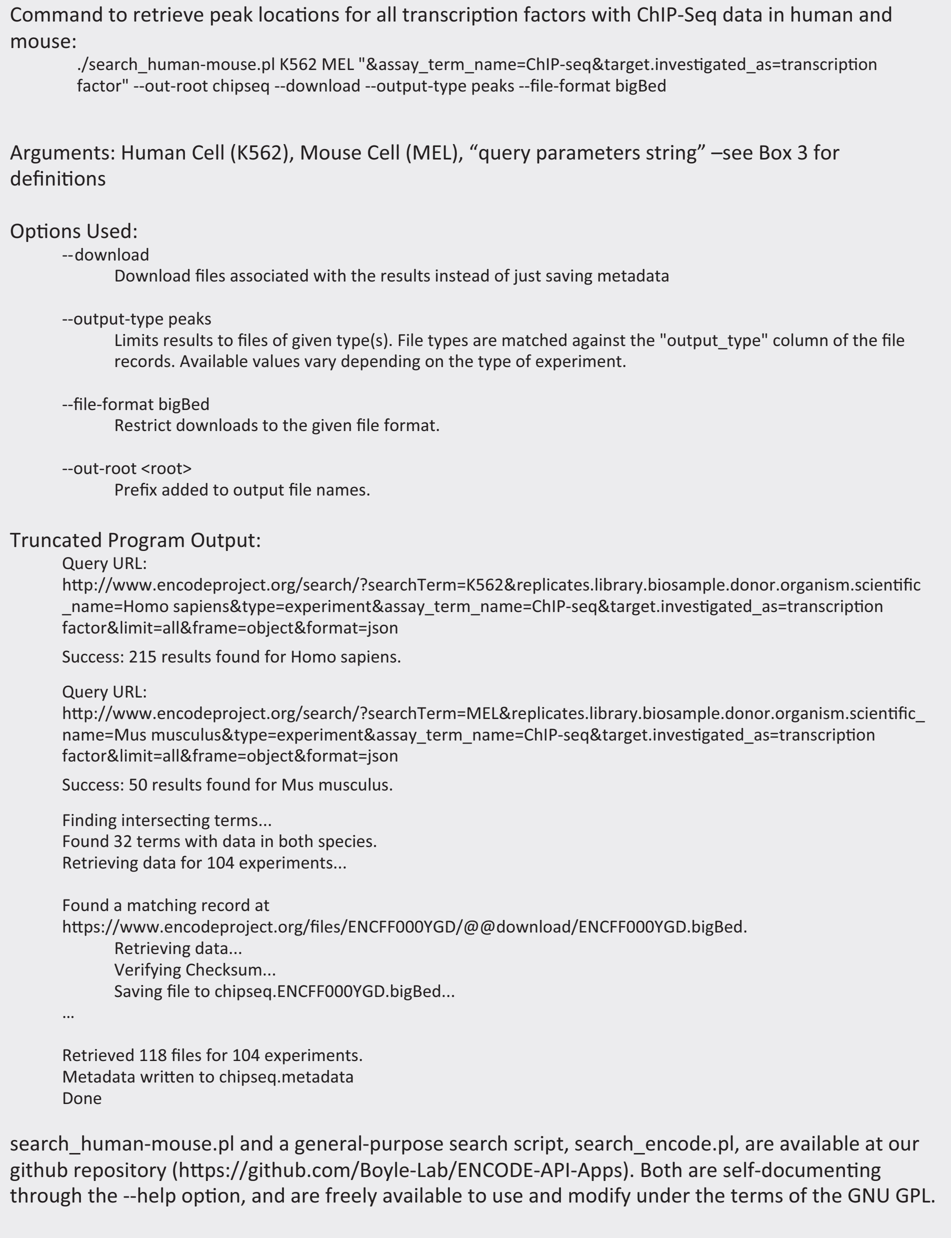
```
1. Process user inputs
   • Check command line to see if we are downloading data files or only retrieving metadata
   • Store specified file and output types, if supplied

2. 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
   • for each 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

3. 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

4. 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

5. 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
```

and a corresponding line of metadata is written to a separate file for later reference. Whereas this process would have been very cumbersome if performed manually, it can be accomplished with a single command using the API (Box 2). It is easy to see how these methods can be extended to more complex analyses, and we encourage readers to make use of our scripts as a starting point. These are available through our github repository (https://github.com/Boyle-Lab/ENCODE-API-Apps), and Box 3 contains a list of helpful API 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 raised 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

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 instead of just saving metadata

--output-type peaks
        Limits results to files of given type(s). File types are matched against the "output_type" column of the file
        records. Available values vary depending on the type of experiment.

--file-format bigBed
        Restrict downloads to the given file format.

--out-root <root>
        Prefix added 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 general-purpose search script, search_encode.pl, are available at our github repository (https://github.com/Boyle-Lab/ENCODE-API-Apps). Both are self-documenting through the --help option, and are freely available to use and modify under the terms of the GNU GPL.

traditionally in the realm of ‘small science’ [30–33]. Indeed, the ENCODE leadership has not always done the best job of embracing a role as a primary data provider, 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. It is a matter for discussion whether being a primary data provider 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

Box 3. API Resources

(Tables I and II)

Table I. 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

Table II. API-Specific Parameters[a,b,c]

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

[a] 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.

[b] ENCODE API help section: www.encodeproject.org/help/rest-api/

[c] ENCODE database schema: www.encodeproject.org/profiles/graph.svg – describes the fields and relationships between tables in the repository database. In many cases, nonstandard search parameters can be built from these using the '.' delimited format used in many of the common parameters.

community, it will continue to do so, yielding further insights that will increase the speed of production, quality, and utility of the data: knowledge that will be of great use to the entire scientific community. Furthermore, because ENCODE data are freely available to everyone, researchers can focus on generating and testing hypotheses rather than on the details of data generation. Its expansion to include the mouse, fly, and worm genomes also increases its impact, and its efforts to extend technological and data-processing advances from human to these species will surely serve the community well. However, as genomics technologies advance, and our understanding of the scientific principles acting upon the genome evolves, so must ENCODE evolve 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. Standardized protocols notwithstanding, batch effects can significantly influence the observed variation between samples, particularly those processed in different laboratories [14,15]. While it is possible to explicitly separate biological and batch effects through careful experimental design, much of the ENCODE data were generated before the importance of such measures was recognized. Even now, they are difficult to fully implement owing to the distributed nature of the project. As a result, biological and batch effects are often confounded, making them difficult to normalize for. Indeed, recent attempts to do so [15,34] have been controversial, and their efficacy remains a matter of intense debate. While the ENCODE leadership has given this considerable attention, it is not yet clear whether measures under discussion by the consortium will be sufficient. Resolving this issue should remain a high priority because identifying effective ways 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 important matter is improving off-the-shelf usability of the data. While ENCODE has done an admirable job of standardizing experimental protocols and sequencing, data-processing pipelines have been far more variable, often leading to significant variation between datasets. Unfortunately, the time and expertise needed to process the raw data in-house represent significant barriers, 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, crucial details, if present at all, are often buried deep within the accompanying PDF documentation files, which appear neither to follow a standardized format nor provide explicit guidelines for what information must be included. Enforcing a uniform format for documentation and making the content directly visible to the API and web portal would allow users to more easily determine when in-house processing is necessary and, when it is needed, assist in choosing the optimal combination of programs and parameters. In the long term, the uniform data processing pipelines currently under development should be applied to both new and existing data to obviate the need for in-house processing.

Finally, while the stated goal of ENCODE 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, it is far from comprehensive. While filling in gaps is an important objective, as new types of assays emerge it becomes harder to prioritize the allocation of ENCODE's limited resources so as to provide relevant data in a timely manner. In short, it is unlikely that ENCODE will ever fully realize its goal (admittedly, a moving target). However, as high-throughput genomics technology becomes 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 broadly useful to consolidate access to such outside datasets using the functionality of the ENCODE web portal and API. Integrating outside datasets that meet the same quality standards as ENCODE data would be highly valuable both in bridging gaps in the ENCODE data and expanding the profile of the repository as a public resource. In particular, greater integration with repositories from complementary consortium projects [e.g., 1000 Genomes, HapMap, and members of the International Human Epigenome Consortium (IHEC), including the Roadmap Epigenome Project] would be a welcome improvement.

Concluding Remarks

Although we have only provided 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 present it in organized, approachable, and flexible ways. From the web portal and API to the range of affiliated and third-party sites offering different views of the ENCODE data, tools are available for many specialized purposes. Furthermore, ENCODE's seamless integration of nonhuman organisms will be invaluable to sorting out common versus species-specific principles of genome biology, and is a key differentiating factor compared to other large genomic consortia. With its focus on breadth of assays, species, and cell types, ENCODE is fundamentally unique, complementing projects such as HapMap, 1000 Genomes, and the IHEC whose approaches are largely depth-based, focused on cataloging the range of variation between human genomes and epigenomes. As our third case study shows, combining these resources enables powerful inquests into genome biology that would not otherwise be possible. Similar to the surge of discoveries following the publication of the human genome, we expect ENCODE, in the broader context of other consortia and individual labs, to yield extraordinary advances in our understanding of how genomes work.

ENCODE has previously been compared to a map: a complete, comprehensive data resource to guide and serve multiple groups over a long period of time [32]. We believe that this is a fitting metaphor for the mission and vision of ENCODE and that, by keeping to this course, it will remain a central resource for many years to come. The ENCODE data are neither perfect nor complete, and there is room for improvement in many areas, but none of the current challenges are insurmountable. On balance, the limitations and controversies surrounding ENCODE are far outweighed by its merits. Efforts to address these challenges are ongoing, and the repository is poised to expand dramatically in the near future, with a concomitant increase in its utility as a resource for functionally annotating the genome. As a testament to its value, the number of publications using ENCODE data continues to grow, providing many important insights on genomic functions, mechanisms, and principles. As the ENCODE resources and datasets continue to grow and improve, their ability to shed light on the many unanswered questions will only increase (See Outstanding Questions).

Outstanding Questions

As massively parallel sequencing becomes more commonplace and affordable, how will ENCODE maintain its relevance? What sets the ENCODE Consortium apart from the community as a whole as a source for primary genomics data?

How will ENCODE overcome the influence of batch effects in interpreting its datasets, particularly in the context of quantitative comparisons between disparate sources?

How can ENCODE improve the usability of its resources, especially to those lacking the resources and/or expertise to perform in-house normalization?

How should ENCODE prioritize its resources between expanding the breadth and depth of the datasets? Would it be appropriate to include non-ENCODE data that meet defined standards in the repository as a way to address the problem of limited resources?

How can ENCODE improve the repository, web portal, and API to improve their usability, flexibility, and power. How can public awareness of these resources be further improved?

Supplemental Information Supplemental information associated with this article can be found, in the online version, at doi:10.1016/j.tig.2016.02.002.

Resources for ENCODE

www.encodeproject.org – main web portal.

www.encodeproject.org/pipelines – descriptions of uniform data processing pipelines.

www.encodeproject.org/software/ – descriptions of software resources used to prepare the ENCODE datasets.

Educational Resources

www.encodeproject.org/help/getting-started/ – overview of ENCODE datasets, formats and access, including navigating with the web site and API.

www.encodeproject.org/help/rest-api/ – ENCODE API tutorial, including example scripts in Python.

www.encodeproject.org/help/file-formats/ – detailed description of file formats used by ENCODE and what they contain.

www.encodeproject.org/tutorials/ – links to tutorials and educational resources, from ENCODE and other sources.

www.encodeproject.org/tutorials/encode-users-meeting-2015/ – archived materials and video from the 2015 ENCODE Users Meeting. Includes practical instructions for using ENCODE resources in a variety of applications and a workshop on using the cloud-based implementations of uniform data-processing pipelines.

www.genome.gov/27553900 – ENCODE tutorials at the National Human Genome Research Institute (NHGRI).

http://genome.ucsc.edu/ENCODE/FAQ/index.html – summary of ENCODE resources at UCSC.

References

1. The ENCODE Project Consortium (2011) *A User's Guide to the Encyclopedia of DNA Elements,* ENCODE
2. Pazin, M.J. (2015) Using the ENCODE resource for functional annotation of genetic variants. *Cold Spring Harb. Protoc.* 2015, 522–536
3. The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816
4. The ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306, 636–640
5. Harrow, J. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774
6. Mouse ENCODE Consortium (2012) An encyclopedia of mouse DNA elements (mouse ENCODE). *Genome Biol.* 13, 418
7. Gerstein, M.B. *et al.* (2010) Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* 330, 1775–1787
8. modENCODE Consortium (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 330, 1787–1797
9. Boyle, A.P. *et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790–1797
10. Cunningham, F. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.* 43, D662–D669
11. Rosenbloom, K.R. *et al.* (2010) ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res.* 38, D620–D625
12. Thomas, D.J. *et al.* (2007) The ENCODE Project at UC Santa Cruz. *Nucleic Acids Res.* 35, D663–D667
13. Yates, A. *et al.* (2015) The Ensembl REST API: Ensembl data for any language. *Bioinformatics* 31, 143–145
14. Li, S. *et al.* (2014) Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* 32, 888–895
15. Gilad, Y. and Mizrahi-Man, O. (2015) A reanalysis of mouse ENCODE comparative gene expression data. *F1000Res* 4, 121
16. Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11, R25
17. Leek, J.T. *et al.* (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11, 733–739
18. Meyer, C.A. and Liu, X.S. (2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.* 15, 709–721
19. Dillies, M.A. *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinformatics* 14, 671–683
20. Taub, M.A. *et al.* (2010) Overcoming bias and systematic errors in next generation sequencing data. *Genome Med.* 2, 87
21. Risso, D. *et al.* (2011) GC-content normalization for RNA-Seq data. *BMC Bioinformatics* 12, 480
22. Leek, J.T. *et al.* (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883
23. Robinson, M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140
24. Johnson, W.E. *et al.* (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 118–127
25. Guenther, C.A. *et al.* (2014) A molecular basis for classic blond hair color in Europeans. *Nat. Genet.* 46, 748–752
26. Ward, L.D. and Kellis, M. (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40, D930–D934
27. Guo, Y. *et al.* (2015) CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* 162, 900–910
28. Nichols, M.H. and Corces, V.G. (2015) A CTCF code for 3D genome architecture. *Cell* 162, 703–705
29. Corradin, O. *et al.* (2013) Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* 24, 1–13
30. Eddy, S.R. (2012) The C-value paradox, junk DNA and ENCODE. *Curr. Biol.* 22, R898–R899
31. Doolittle, W.F. (2013) Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci. U.S.A.* 110, 5294–5300
32. Eddy, S.R. (2013) The ENCODE project: missteps overshadowing a success. *Curr. Biol.* 23, R259–R261
33. Graur, D. *et al.* (2013) On the immortality of television sets: 'function' in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* 5, 578–590
34. Lin, S. *et al.* (2014) Comparison of the transcriptional landscapes between human and mouse tissues. *Proc. Natl. Acad. Sci. U.S.A.* 111, 17224–17229