Topic Introduction

Using the ENCODE Resource for Functional Annotation of Genetic Variants

Michael J. Pazin¹

*Division of Genome Sciences, National Human Genome Research Institute, National Institutes of Health,
Bethesda, Maryland 20892*

This article illustrates the use of the Encyclopedia of DNA Elements (ENCODE) resource to generate or refine hypotheses from genomic data on disease and other phenotypic traits. First, the goals and history of ENCODE and related epigenomics projects are reviewed. Second, the rationale for ENCODE and the major data types used by ENCODE are briefly described, as are some standard heuristics for their interpretation. Third, the use of the ENCODE resource is examined. Standard use cases for ENCODE, accessing the ENCODE resource, and accessing data from related projects are discussed. Although the focus of this article is the use of ENCODE data, some of the same approaches can be used with data from other projects.

GOALS AND HISTORY OF ENCODE

The primary goals of ENCODE, the Encyclopedia of DNA Elements, are (i) to create a comprehensive catalog of candidate functional elements in the genome, and (ii) to make that catalog freely available as a community resource for all biologists. ENCODE resources can be accessed from the ENCODE portal (<http://www.encodeproject.org>) and at other URLs (Box 1). ENCODE data (transcription, transcription factor binding, histone modifications, DNase hypersensitivity, DNA methylation, DNA–DNA interactions, and RNA–protein interactions) are rapidly released to the public before publication, following the precedent of the human genome project. External users may freely download, analyze and publish results based on any ENCODE data (without any embargo or restrictions) as soon as they are released. ENCODE is focused on the human genome, though ~20% of the data collected annotate the mouse genome. The fly and worm genomes were the focus of the model organism (mod) modENCODE project. The catalogs, or maps, of candidate elements are intended to complement ongoing efforts to understand the functions resident in the genome, rather than to replace those individual efforts. At this time, ENCODE has released approximately 3000 human experiments, each containing at least two replicates, examining approximately 200 cell types (cell lines, primary cells, cells differentiated in culture, and explants), and approximately 900 mouse experiments, each containing at least two replicates, in more than 100 cell lines, primary cells, and explants. To date, ENCODE human and mouse data have appeared in approximately 650 papers published by researchers outside of ENCODE, and modENCODE data have appeared in approximately 150 papers by researchers outside of modENCODE (https://www.encodeproject.org/search/?type=publication&published_by=community).

ENCODE was launched in 2003 with a pilot project to survey 1% of the human genome (Consortium 2004). Major findings from this microarray-based pilot phase were published in several papers in 2007 (e.g., Birney et al. 2007). Based on the success of the pilot project, a genome-wide production

¹Correspondence: pazinm@mail.nih.gov

Cite this introduction as *Cold Spring Harb Protoc*; doi:10.1101/pdb.top084988

BOX 1. INTERNET RESOURCES FOR ENCODE

ENCODE Portal:

<http://www.encodeproject.org>

The ENCODE Portal has resources for searching, downloading, and visualizing ENCODE mouse and human data. The portal also has data summaries, an experiment list, consortium publications, community publications using ENCODE data, software tools, quality metrics, and data standards.

ENCODE project pages, NHGRI:

<http://www.genome.gov/encode/>

Tutorials on using the ENCODE resource:

<http://www.genome.gov/27553900>

<https://www.encodeproject.org/tutorials>

Automated mining of ENCODE data:

<http://www.broadinstitute.org/mammals/haploreg/haploreg.php>

<http://regulome.stanford.edu>

<http://regulome.stanford.edu/GWAS>

http://www.genome.gov/Pages/Research/ENCODE/ASHG_2013_Using_HaploReg_RegulomeDB_to_Mine_ENCODE_Data.pdf

Linkage between genes and regulatory elements:

<http://dnase.genome.duke.edu>

http://www.genome.gov/Pages/Research/ENCODE/ASHGASHG_2013_Predicting_Target_Genes_%20For%20Distal_Regulatory_Region.pdf

http://www.genome.gov/Pages/Research/ENCODE/ASHG_2013_Predicting_Distal_Regulatory_Regions_For_A_Gene.pdf

Some info on visualizing ENCODE data at a region of interest:

<https://www.encodeproject.org/encyclopedia/visualize>

http://www.genome.gov/Pages/Research/ENCODE/ASHG_2013_Viewing_ENCODE_Composite_Tracks_Locus_Of_Interest.pdf

How labs within and outside of ENCODE use the data in publications:

<https://www.encodeproject.org/publications>

https://www.encodeproject.org/search/?type=publication&published_by=community

https://www.encodeproject.org/search/?type=publication&published_by=ENCODE&published_by=mouseENCODE

ENCODE and Roadmap Epigenomics data:

<http://epigenomegateway.wustl.edu>

<http://www.encode-roadmap.org>

ENCODE mailing list:

<https://mailman.stanford.edu/mailman/listinfo/encode-announce>

phase using massively parallel sequencing focused on the human genome was launched in 2007, and production efforts focused on the mouse were begun in 2009 (Stamatoyannopoulos et al. 2012). In addition, projects to study the fly and worm genomes were launched, to improve and to supplement annotations of the genomes of these important model organisms, as well as to assist in interpretation of the human genome (Celniker et al. 2009; Gerstein et al. 2010; Roy et al. 2010; Graveley et al. 2011; Kharchenko et al. 2011; Negre et al. 2011). A user's guide to ENCODE (Consortium 2011) and recent findings on the human genome (e.g., Consortium 2012) have been published. The ENCODE and modENCODE projects are funded by the National Human Genome Research Institute (NHGRI), one of the components of the National Institutes of Health (NIH, U.S.).

RELATED EPIGENOMICS PROJECTS

REMC, the Roadmap Epigenomics Mapping Centers project, is funded by the Common Fund of the NIH (U.S.). The REMC goal is to generate reference epigenomic maps for “normal” human cells/tissues. Most samples have RNA-seq, DNase-seq, DNase-seq, and histone modification data.

IHEC, the International Human Epigenome Consortium, is an international project. The IHEC goal is to coordinate production of 1000 human epigenome maps for cellular states relevant to health and disease. Complete epigenomes include at least mRNA-seq, DNase-seq/WGBS, and six histone modifications. IHEC projects generally include samples from healthy people and targeted diseases. IHEC projects include BLUEPRINT (European Union), CEEHRC Epigenomic Platform Program (Canada), CREST (Japan), and DEEP (Germany). In addition, REMC (Roadmap Epigenomics) is a member of IHEC, and ENCODE is an associate member of IHEC. ENCODE, REMC, and IHEC are coordinating metadata and ontologies to improve the interoperability of the data from the projects.

GTEx, the Genotype-Tissue Expression project, is funded by the Common Fund of the NIH (U.S.). The goal of GTEx is study human gene expression across individuals, and across multiple tissues. Most samples have RNA-seq and genotype. Having completed a feasibility pilot phase, GTEx has recently entered into a production phase.

URLs for descriptions of these projects and data distribution are included (Box 2).

RATIONALE FOR ENCODE

ENCODE directly addresses a need identified during the human genome project. The goal of the human genome project was to obtain a complete human genome sequence, and make that sequence available as a public resource. One of the first efforts to make practical use of this sequence was the GENCODE project (an ENCODE component, the Encyclopedia of genes and gene variants). GENCODE has produced a widely used annotation for protein-coding genes (Harrow et al. 2012). Twenty-one annotated versions of the human genome have been developed with both protein and nonprotein coding genes now available to the scientific community (Harrow et al. 2012). The GENCODE v21 human annotations indicate there are 19,881 protein-coding genes and 25,411 non-coding genes.

The genetic code, while very powerful, is only useful for interpreting protein-coding regions (~1.2% of the human genome). However, the genome also contains regulatory regions, and as we do not have a similarly powerful regulatory code to call on, it is beyond our grasp today to even identify all of the functional (or nonfunctional) regions of the genome by sequence alone, let alone to determine when, where, and how they act. Regulatory regions often appear to control distal genes, which are not necessarily the nearest gene, and often function in particular cell fates or cell states. Moreover, a particular regulatory region can apparently control more than one gene. ENCODE is one effort to fill this gap in our knowledge using unbiased experimentation.

Another area in which ENCODE is playing a role in assisting the scientific community is in the area of helping to investigate the genetic components of human disease and traits. Genetic variants that are statistically associated with traits of interest are identified using genome-wide association studies (GWAS), exome sequencing, and whole genome sequencing. These powerful studies generate testable hypotheses that can be explored in subsequent studies. The ultimate goal of these lines of experimentation is often to identify better predictions of disease risk, better diagnostics, and better therapeutics.

Unfortunately, statistical association studies have limitations that are fundamental, and well known. First, a variant associated with a trait or disease is not necessarily the causal variant, in part because of linkage disequilibrium and in part because most studies do not interrogate the whole genome. Second, the gene that is affected by a variant can be difficult to identify, as regulation can occur at large distances, and target multiple genes. Third, the cell fate(s) and cell state(s) affected by germline variants can be difficult to surmise. This information is important given that both complex and Mendelian disorders can involve multiple cell types. Finally, it is often not straightforward to determine how causal variants alter cells and organisms, as one must consider changes in transcription, splicing, RNA stability, translation, and protein function, among others. It is important to note that >90% of GWAS findings lie outside of protein coding regions (Thurman et al. 2012), and >80% of the findings for signatures of recent adaptive evolution lie outside of protein-coding regions (Jones et al. 2012; Fraser 2013; Grossman

BOX 2. INTERNET RESOURCES FOR OTHER EPIGENOMICS PROJECTS

Roadmap Epigenomics Mapping Centers (REMC; USA, NIH Common Fund):

Project description:

<http://commonfund.nih.gov/epigenomics/index>

Data visualization:

<http://www.roadmapepigenomics.org/>

<http://www.genboree.org/epigenomeatlas/>

<http://encode-roadmap.org>

<http://epigenomegateway.wustl.edu>

(Also Track hub on UCSC, Ensembl, and ENCODE portal)

International Human Epigenome Consortium (IHEC; Multinational):

Project descriptions:

<http://www.ihec-epigenomes.org/>

<http://ihec-epigenomes.org/research/projects/>

Data distribution:

<http://epigenomesportal.ca/ihec/>

<http://ihec-epigenomes.org/outcomes/data-distribution/>

BLUEPRINT (European Union, IHEC member):

Project description:

<http://www.blueprint-epigenome.eu>

Data visualization:

<http://www.blueprint-epigenome.eu/index.cfm?p=7DEC5013-D416-AFDB-BFEABB82782623BB>

(Also Track hub on UCSC, Ensembl, and ENCODE portal)

CEEHRC Epigenomic Platform Program (Canada, IHEC member):

Project description:

<http://www.epigenomes.ca>

<http://ihec-epigenomes.org/research/projects/epigenomic-platform-program/>

Data distribution:

<http://epigenomesportal.ca/ihec/>

(Also Track hub on UCSC, Ensembl, and ENCODE portal)

CREST (Japan, IHEC member):

Project description:

<http://crest-ihec.jp/english/index.html>

Data distribution:

<http://crest-ihec.jp/english/database/index.html>

DEEP (Germany, IHEC member):

Project description:

<http://www.deutsches-epigenom-programm.de>

GTEx (Genotype-Tissue Expression, US, NIH Common Fund):

GTEx portal:

<http://www.gtexportal.org/>

GTEx project site:

<http://commonfund.nih.gov/GTEx>

et al. 2013). Individual examples, such as Fragile X, ALS, lactose tolerance, and polydactyly, also indicate that variation in noncoding DNA makes important contributions to human traits and disease, including Mendelian and severe traits (Lettice et al. 2002; Penagarikano et al. 2007; Tishkoff et al. 2007; Kleinjan and Lettice 2008; Herdewyn et al. 2012). When we fail to consider the role of noncoding DNA, we are likely ignoring the vast majority of findings, which will likely diminish our power to understand human diseases and traits. The contributions of ENCODE in providing both gene and regulatory annotations for these noncoding regions has recently proven useful (https://www.encodeproject.org/search/?type=publication&published_by=community&categories=human+disease).

Epigenomic and transcriptomic experimentation in general, and the ENCODE resource in particular, can be used to generate and refine hypotheses to allow researchers to move from statistical

associations to functional connections (Hardison 2012; Hardison and Taylor 2012; Ward and Kellis 2012c; Edwards et al. 2013). This information is especially powerful with noncoding variants (Maurano et al. 2012), which is fortunate because most GWAS findings are noncoding, and we lack a regulatory code with predictive power similar to the genetic code.

ENCODE DATA TYPES

The major ENCODE approach to identifying candidate functional elements is to use biochemical assays that are capable of surveying the entire genome. Because functional elements may be cell-type-specific, these biochemical assays are performed in many cell types, to identify the largest and most diverse fraction of the functional elements that is practically achievable. In addition, ENCODE is, to a more limited extent, using the approach of comparative sequence analysis to identify and characterize candidate functional elements. These two fundamentally different approaches have different strengths and weaknesses; the same can be said of genetic approaches. As was the case before ENCODE, the findings from these different approaches do not always agree, so resolving these differences continues to be an active area of research (Nardone et al. 2004; Borneman et al. 2007; Odom et al. 2007; Noonan and McCallion 2010; Schmidt et al. 2010; Biggin 2011; Fisher et al. 2012; MacArthur et al. 2012).

Conceptually, ENCODE is attempting to identify genes, transcribed regions, and transcripts, as well as regulatory regions in DNA and RNA (Fig. 1). The biochemical assays used by ENCODE have been widely used by other scientists. Indeed many of the underlying assays, such as DNase I hypersensitivity site (DHS) mapping (Wu 1980; Gross and Garrard 1988), and chromatin immunoprecipitation (ChIP) (Gilmour and Lis 1984; Solomon et al. 1988; Ren et al. 2000; Iyer et al. 2001; Litt et al. 2001; Johnson and Bresnick 2002) were initially developed by individual laboratories and later were adapted to genome-wide assays. Many of these assays have long been used in the study of gene regulation (Gross and Garrard 1988; Johnson and Bresnick 2002; Weinmann and Farnham 2002; Nardone et al. 2004; Li et al. 2006; Ozsolak et al. 2008; Visel et al. 2009; Creighton et al. 2010; Rada-Iglesias et al. 2011). As these

FIGURE 1. Scheme of ENCODE data types. ENCODE identifies candidate genes (transcribed regions, and transcripts) as well as regulatory elements in DNA and RNA (*bottom*). ENCODE researchers use a variety of biochemical assays (rounded rectangles in *middle*) to identify features associated with DNA and RNA (drawing at *top*).

methods are used and refined, the new knowledge results in new interpretations of findings from these assays; however, the conceptual interpretation remains quite stable, as findings accumulate from many independent laboratories over many years (reviewed in Rivera and Ren 2013).

RNA data are used to identify both protein-coding genes, and also noncoding genes such as miRNAs and lncRNAs (reviewed in Djebali et al. 2012). The prevalence of nonpolyadenylated poly(A)[−] RNA is increasingly recognized as potential functional elements based on recent studies, including ENCODE (Kapranov et al. 2007; Tilgner et al. 2012; Livyatan et al. 2013). Transcriptional start sites can be identified using assays such as CAGE and RAMPAGE, specifically designed for this purpose (Shiraki et al. 2003; Batut et al. 2013). Active enhancers can sometimes be identified by the production of poly(A)[−] RNAs known as eRNA that are on average shorter than most protein coding and noncoding transcripts. In addition, transcribed regions can be assembled into previously annotated and unannotated candidate transcripts (Harrow et al. 2012).

DNase hypersensitivity has long been used to identify candidate regulatory elements (Weintraub and Groudine 1976; Gross and Garrard 1988; Fraser et al. 1993; Bell et al. 2011b; Thurman et al. 2012). Binding of regulatory proteins often perturbs chromatin structure, leading to DNase hypersensitivity (Pazin et al. 1994, 1996; Henikoff and Shilatifard 2011). A recent genome-wide survey revealed that expression quantitative trait loci (eQTLs) are frequently associated with genotype-dependent changes in DNase (dsQTLs, DNase I sensitivity quantitative trait loci), suggesting DHSs are frequently regulatory regions (Degner et al. 2012). Regulatory regions identified by DHSs include enhancers and promoters, as well as locus control regions and silencers (Fraser et al. 1993; Diaz et al. 1994; Sawada et al. 1994; Siu et al. 1994; Festenstein et al. 1996). Recent studies have found DHS to be useful for annotating GWAS findings (Consortium 2012; Maurano et al. 2012; Pickrell 2013). Cell types clustered by DHS fall into biologically meaningful groups by developmental lineage and age, reflecting the information contained in DHS signatures (Stergachis et al. 2013).

Histone modifications (acetylation and methylation) have long been considered as surrogate marks for gene activity (Allfrey et al. 1964; Allis et al. 2007), and have been studied by chromatin fractionation, staining of chromosomes, and ChIP (Levinger and Varshavsky 1980; Hebbes et al. 1988; Lin et al. 1989; Turner 1989; Hebbes et al. 1994). Many histone modifications and histone variants are known, and a number have relatively well-accepted correlations with gene expression (e.g., Campos and Reinberg 2009; Ernst and Kellis 2010; Huff et al. 2010; Bell et al. 2011b; Ernst et al. 2011; Suganuma and Workman 2011; Consortium 2012). For example, promoters are enriched for H3K4me3, while enhancers are enriched for H3K4me1 (Heintzman et al. 2007). Active enhancers and promoters are often enriched for H3K27ac (Creighton et al. 2010; Rada-Iglesias et al. 2011). Repressed regions of the genome are enriched for H3K27me3 and H3K9me3; the latter is found in constitutive heterochromatin, while the former is found in regions repressed during development by the polycomb system. Transcribed regions are enriched for H3K36me3 and H3K79me3; near the first intron, there appears to be a transition from H3K79 to H3K36 (Huff et al. 2010; Consortium 2012). Combinations of histone marks have been used to identify enhancers, promoters, and transcribed regions (Heintzman et al. 2007, 2009; Ernst and Kellis 2010; Ernst et al. 2011; Consortium 2012). Global analysis of histone modifications over many cell types confirms correlation with gene expression (Consortium 2012).

Transcription factor binding has been used for some time to identify candidate regulatory elements (Johnson and Bresnick 2002; Weinmann and Farnham 2002; Gerstein et al. 2012; Maston et al. 2012; Wang et al. 2012; Yip et al. 2012). ENCODE determines transcription factor binding regions using ChIP, DNase footprinting, and DNA sequence analysis, and has also analyzed the networks formed by these transcription factor binding sites (Gerstein et al. 2012; Neph et al. 2012a, 2012b; Wang et al. 2012; Yip et al. 2012; Kheradpour and Kellis 2013). ENCODE has grouped measurements of over 100 factors, including sequence-specific transcription factors, RNA polymerase subunits, general transcription factors, and chromatin remodeling enzymes, within this data type. Global analysis of transcription factor binding is highly correlated with gene expression, and GWAS findings are enriched in binding sites, confirming the importance of the identification of transcription factor binding sites (Consortium 2012).

DNA methylation at promoters and enhancers is inversely correlated with gene expression, while DNA methylation at gene bodies is correlated with gene expression (Varley et al. 2013). DNA methylation is inversely correlated with open chromatin, transcription factor binding, and histone modifications found in active chromatin (Bell et al. 2011a; Stadler et al. 2011). Overlap in variants that alter DNA methylation and variants that alter gene expression has been detected, and variation in DNA methylation has been found to possess a strong genetic component (Bell et al. 2011a). Cell types clustered by DNA methylation fall into biologically meaningful groups, revealing lineage and cancer signatures, reflecting the information contained in DNA methylation (Varley et al. 2013).

To date, ENCODE has not focused on characterizing long-range interactions of regulatory regions. However, ENCODE has collected limited amounts of both 5C and ChIA-PET data, and plans to continue to do so (Li et al. 2012; Sanyal et al. 2012). Both data types identify interactions between regulatory elements, which are candidate functional interactions. 5C identifies DNA–DNA interactions, while ChIA-PET identifies DNA–DNA associations bound by a protein of interest. To date, the primary utility of this data has been to validate predictions of enhancer–promoter linkage, to predict the genes that are targets of regulatory elements. This work appears to validate the approach, in that there is good correlation between gene expression, enhancer–promoter interaction, and production of eRNA. This work also finds that the majority of looping interactions are not to the most proximal gene (Sanyal et al. 2012).

A variety of candidate functional elements can be identified by sequence conservation, including protein-coding and noncoding genes, as well as regulatory regions (e.g., Nardone et al. 2004). Sequence comparison has revealed evidence of selection for 3%–15% of the human genome; most of this DNA does not appear to code for protein (reviewed in Ponting and Hardison 2011). The ENCODE estimate of constraint (3%–10%) is consistent with estimates by others (3%–15%) (Ponting and Hardison 2011; Ward and Kellis 2012a). ENCODE estimates of the amount of human regulatory DNA (5%–20%, from DNase footprints to DHS + TF-ChIP) (Consortium 2012) agrees well with estimates of constraint. ENCODE has also compared biochemical and sequence-comparison based approaches to identification of functional elements (Birney et al. 2007; Consortium 2012).

USE CASES FOR THE ENCODE RESOURCE

ENCODE data have been used to suggest or refine hypotheses about the role of genetic variation in phenotypic variation (including disease). Publications by consortium members (https://www.encodeproject.org/search/?type=publication&published_by=ENCODE), as well as by researchers outside of ENCODE (https://www.encodeproject.org/search/?type=publication&published_by=community) illustrate a number of approaches for using the ENCODE resource in the study of human disease, basic biology (including gene regulation), and methods development. This section will present some common use cases for ENCODE data. Although the focus here is on interpreting genetic variation data, essentially the same approaches can be used with the ENCODE resource or with data from other projects, with appropriate modification, to interpret epigenomic and gene regulation data (Rakyan et al. 2011; Ng et al. 2012). Such approaches could allow investigators to use genomic methods to study environmental and stochastic processes, in addition to genetic processes.

Use Case: Prediction of Causal Variants

Genetic association studies identify statistical association between a trait of interest and a genetic variant (referred to here as the “tag variant”). While it is possible the tag variant causes the phenotypic change of interest, for at least two fundamentally different reasons it is also possible that another variant (referred to here as the “causal variant”) is responsible. First, this can happen when multiple variants are correlated through linkage disequilibrium; the significance of the association can be similar enough for these variants to make it difficult to distinguish them. Second, the functional variant may not even have been tested in the study; GWAS arrays interrogate a subset of the known

common variants, and exome sequencing covers a limited fraction of the genome. It is also possible for a tag variant in a protein-coding region to mark a causal variant in a regulatory region (Herdewyn et al. 2012). While there is no one best approach to identify the causal variant, epigenomic and transcriptomic data can be used to refine hypotheses (Schaub et al. 2012).

One approach is to use semiautomated tools to analyze ENCODE data. RegulomeDB (Boyle et al. 2012) and HaploReg (Ward and Kellis 2012b) mine the data in ENCODE, as well as other resources, and return information on candidate functional variants. RegulomeDB accepts SNP IDs, single nucleotides, and chromosomal regions as inputs (Fig. 2). RegulomeDB provides a score for each variant queried, with lower scores indicating more supporting evidence; for example, eQTLs score in the range 1a–1f, strong evidence of binding without an eQTL scores 2a–2c, continuing out to 6, followed by a score of no evidence. Users can click on the provided hyperlinks to examine the underlying evidence (which includes TF-ChIP signals, DNase peaks, DNase footprints, and predicted DNA sequence motifs for TFs) or to a genome browser view of the region. RegulomeDB also functions as a database, presenting a list of traits and conditions as hyperlinks, which can be selected to display associated variants, the publications that identified them, scores for variants within LD of the query, and links to RegulomeDB annotations. HaploReg accepts SNP IDs, chromosomal regions, or published GWAS studies as inputs, and also a user-configurable definition of linkage disequilibrium (Fig. 3). HaploReg returns all variants within the query (as well as variants within LD), and indicates what evidence (if any) supports a functional role for each variant. While HaploReg was originally designed to mine ENCODE data, it now also mines data from the NIH Common Fund project Roadmap Epigenomic Mapping Centers (REMC). As RegulomeDB and HaploReg use somewhat different approaches, it is probably beneficial to use both for hypothesis generation.

Another approach is to manually inspect the region of interest to determine what transcriptomic and epigenomic annotations are present. Starting from the ENCODE portal annotations at <https://www.encodeproject.org/data/annotations>, one can view ENCODE annotations (see <https://www>.

FIGURE 2. Annotating potential regulatory variants using RegulomeDB. RegulomeDB accepts a variety of inputs, including SNP IDs and genomic coordinates (red arrow 1). RegulomeDB returns a score for each feature in the input region (red arrow 2); scores range from 1 (strongest evidence for regulatory potential) to 6 and no data (weakest evidence). Clicking on the score reveals the details of the evidence; the data types, cell fate/cell state, and the source of the data. For example, the indicated row (red arrow 3) displays an ENCODE DHS in lymphoid cells. Clicking on the browser link next to the score provides a graphical visualization of the evidence; for example, ENCODE transcription factor binding is present (red arrow 4).

FIGURE 3. Annotating potential regulatory variants using HaploReg. HaploReg accepts SNP IDs and genomic coordinates as inputs (red arrow 1). HaploReg returns evidence for regulatory protein binding (mouse over to see the protein names), chromatin structure (mouse over to see the cell types with DNase hypersensitivity), the chromatin state of the region, and putative transcription factor binding motifs that are altered by the variant (row marked by red arrow 2). SNPs in LD with the query are also reported in the rows above and below.

encodeproject.org/encyclopedia/visualize) of distal DHS and H3K27ac sites (which can be considered as candidate enhancers), and proximal sites (which are candidate promoters or enhancers). The ENCODE portal offers browsing of ENCODE data based on assay types and biological samples, via faceted browsing, to find data to visualize and download (Fig. 4). Starting from the UCSC browser (<https://genome.ucsc.edu/encode/>), the region of interest can be viewed along with ENCODE tracks. For example, one can start with a display of transcribed regions, H3K4me1 (an enhancer mark), H3K4me3 (a promoter mark), H3K27ac (a mark for active promoters and enhancers), DNase (an active chromatin mark frequently found at regulatory regions) and transcription factor ChIP data (frequently found at regulatory regions), as recently described (Mortlock and Pregizer 2012). Using the sessions feature of the UCSC browser (Meyer et al. 2013), users can save their favorite configuration(s) to load them for future use, or to share with colleagues (instructions at http://www.genome.gov/Pages/Research/ENCODE/2012_09_18_ENCODE_Sessions.pdf). One can also use this approach to examine particular tracks of interest, such as comparing B cells, T helper subsets, and macrophages. One strength of manual visualization is the user can simultaneously visualize annotations from multiple projects, such as ENCODE, REMC, and IHEC, by loading the desired track hubs (instructions at http://www.genome.gov/Pages/Research/ENCODE/2012_09_18_ENCODE_Sessions.pdf).

Use Case: Prediction of Target Genes

Even if one could be certain they had identified a functional variant, the gene that is affected by the variant may not be immediately obvious, for at least two fundamentally different reasons. First, it has been known for some time that regulatory regions can regulate distal genes, multiple genes, and the

FIGURE 4. Browsing of ENCODE data using the ENCODE Portal. The ENCODE Portal (<http://encodeproject.org>) allows users to select subsets of ENCODE data, for example by data type (red arrow 1), organism (red arrow 2), or organ (red arrow 3). The identified data can be visualized or downloaded.

target gene(s) is not necessarily the nearest transcription start site (TSS) (Forrester et al. 1987; Fraser et al. 1993; Bedell et al. 1995; Mohrs et al. 2001; Lee et al. 2003). Moreover, distal regulatory regions are known to contribute to human phenotypes and disease, including traits segregating in Mendelian fashion (Lettice et al. 2002; Loots et al. 2005; Kleinjan and Lettice 2008; Noonan and McCallion 2010). Data from ENCODE add support to the ideas that distal regulatory regions, and regulatory regions that target multiple genes, may not be rare (Thurman et al. 2012). If this is not taken into account, one might identify the incorrect target for developing therapeutics, or for pathway analysis (Lettice et al. 2002; Kleinjan and Lettice 2008; Noonan and McCallion 2010; Davison et al. 2012; Maurano et al. 2012). Second, the target gene could be a nonprotein coding RNA, and the annotation of these regions is still being developed.

Currently, there are no high-throughput methods to determine the functional connectivity between enhancers and promoters; this is typically done through intensive study of single loci. However, ENCODE has produced DHS maps from over 100 samples, and many DHS are cell-type-specific. The Regulatory Elements Database (Sheffield et al. 2013) is an Internet-based tool (<http://dnase.genome.duke.edu>) that can be used to predict the target gene for DHSs of interest, or to predict the DHSs that regulate a gene of interest (Fig. 5). The predictions are based on correlations between the cell specificity of gene expression and the cell specificity of DHSs, which are candidate regulatory regions. As both positive and negative correlations are reported, this tool may also predict the linkage between genes and silencers, the repressive version of enhancers. In an earlier effort, the correlation between the cell specificity of distal DHS and the cell specificity of promoter DHS (a surrogate of gene expression) was tabulated (Thurman et al. 2012). Supplementary Table 7 from Thurman et al. (2012) can be used to predict the linkage between genes and regulatory regions. Using one-line Unix commands, this table can be sorted by DHS start coordinate (sort -k 5,5 -k 6,6n path/filename > path/newfilename) or by the name of linked genes (sort -k 4,4 path/filename > path/newfilename). As these two resources can identify different linkages, likely because of the difference in approach, it is useful to use both for hypothesis generation. An important feature of both of these tools is they integrate data from large numbers of cell types to provide specificity.

FIGURE 5. Predicting genes associated with a regulatory element using the Regulatory Elements Database. An intuitive interface facilitates identification of known DNaseI-hypersensitive sites (DHS) within a genomic region of interest, prediction of the target gene for DHS of interest, and prediction of the DHS that regulates a gene of interest, as well as other tasks (red arrow 1). After finding a DHS within a coordinate range (red arrow 2), one can see this DHS is positively correlated with the IL13 gene, predicting this region activates IL13 (red arrow 3). The DHS profile across cell types reveal this element is active in Th2 cells, and inactive in most of the other cell types (red arrow 4), suggesting variants in this region could be important in a particular type of T cell.

ENCODE RNA data can also be used in predicting target genes. One can identify the transcribed regions (protein coding and noncoding) near a variant of interest, and determine the cell types they are expressed in Hu et al. (2011). If the relevant cell type(s) are known, it may be possible to refine the list of candidate genes.

Use Case: Prediction of the Cell Fate/Cell State where Elements (and Variants) Alter Function

Identification of a tag variant, causal variant, or target gene does not always reveal the pathological cell type, for at least three fundamentally different reasons. First, some diseases, such as cardiovascular disease, are known to involve and affect multiple cell types. Genetic association findings in such diseases cannot always be immediately assigned to the relevant cell type(s). Second, the defect may not be intrinsic to the cell type that displays obvious pathology, and may not initially be recognized as relevant to the trait or disease. For example, a rare human immunodeficiency appears to be caused by the same genetic defect found in *nude* mice (Frank et al. 1999). In the mouse model, it is well known that the immunodeficiency is intrinsic to the epithelial cells that form the niche for developing T cells, rather than the affected lymphocytes (Nehls et al. 1996). A third potential confounder is that frequently disease etiology is incompletely known; for example, it was not always appreciated that Type 1 Diabetes and Crohns Disease were autoimmune disorders.

The previously described tools for identifying causal variants (HaploReg, and RegulomeDB) and linkage between regulatory elements and genes (Regulatory Elements Database) report the cell type where the evidence for active regulatory elements was found. This information can be used to generate hypotheses about the cell type affected by the genetic variants. This information is even more powerful when one starts with a trait or disease with substantial information on the relevant cell types. In this case, one may be attempting to distinguish between very different cell types (e.g., neurons, muscle, or lymphocytes) in which case ENCODE annotations in a cell type *related* to one of the disease-relevant cell types may be sufficient to predict the cell type where the causal variant functions.

Manual inspection of ENCODE tracks at a locus of interest also reports information about the cell type. ENCODE composite tracks, which combine data from multiple cell types, can be expanded to reveal the contributions from individual cell types. One can select individual tracks from particular cell types, either to broadly examine the signal over widely different cell types, or closely related cell types.

RNA data from ENCODE can be used to predict affected cell types. One can examine the cell specificity of expression for genes (protein coding and noncoding) in the vicinity of a variant of interest (Hu et al. 2011). One can also determine the expression patterns of genes that are predicted to be linked to a variant of interest. Finally, if the variant is a candidate regulatory element, one can determine whether cell-type-specific eRNAs are produced by the locus.

A more rigorous way to predict cell type is to calculate enrichment of variants (associated with a trait or disease of interest) in regulatory elements, by cell type. Enrichment of variants within DHSs and histone modifications sites from ENCODE and the Roadmap Epigenomics projects have been used to predict known and plausible cell types in test cases (Maurano et al. 2012; Pickrell 2013; Trynka et al. 2013). This approach can also be used to increase the number of testable variants associated with a trait or disorder, moving beyond statistical thresholds to biological thresholds. One can initially apply a statistical threshold to identify relevant variants, then determine enrichment by cell type, next relax the threshold as far as possible without losing the predicted cell specificity, and then test this larger list of variants (Maurano et al. 2012). This approach could be especially useful in settings where the capacity to test variants substantially exceeds the number of predicted variants.

SUMMARY

Development of the ENCODE resource is an ongoing effort, informed by community feedback, and the ever-changing landscape of genomics technology. Clearly a broader and deeper sampling of cell fate/cell state space is an important need. During the current phase of the project, ENCODE plans to examine more cell types, with an emphasis on explants and primary cells, and an emphasis on increasing the number of samples with many ENCODE assays. Users have found value in both explants and purified cells (though not always the same users), so ENCODE will likely produce data on both sides of this compromise between better-defined samples and more-physiologic samples. However, the cell fate/cell state space is enormous. The difficulty of identifying functional connections between regulatory regions and genes, which was recognized before ENCODE, has become more apparent with the wealth of new data from a variety of sources. ENCODE is continuing data collection on physical interactions between candidate regulatory elements, as a high-throughput surrogate measurement of functional interactions. ENCODE plans to characterize at least 500 transcription factors, a significant increase over the current total of ~100; again, considering there are in the order of 2000 transcription factors, and they appear to bind different motif instances in different cell states and fates, the space is enormous. Given the growing recognition of the importance of regulation at the RNA level, ENCODE is expanding measurements of RNA binding proteins from pilot scale to production scale, with a goal of characterizing about 200 RNA binding proteins. These efforts will consider proteins that bind mRNA, and regulatory proteins that likely use noncoding RNA as an adapter to bind to targets. ENCODE is working with our partners in other epigenomics projects to increase the utility of the data from all the projects. We are working together to coordinate metadata and ontologies, for example, in an attempt to make it easier to use the data from different projects in the same way, or to bring together data from different projects. Finally, ENCODE has relaxed the data use policy for external users. As soon as ENCODE data are publicly released, they may be used in publications and presentations. Previously there had been a moratorium on the global use of newly released data sets (for either 9 mo or the first consortium publication). However, this created some confusion over what data were subject to what restrictions; we have eliminated this moratorium to simplify this aspect of using ENCODE data. We continue to ask that users cite ENCODE, as would be the case for any other scientific source used.

ACKNOWLEDGMENTS

Thanks to these individuals for their helpful comments on draft versions of this manuscript: Brad Bernstein, Guillaume Bourque, Elise Feingold, Tom Gingeras, Eurie Hong, Robert Klein, Jonathan Pritchard, Bing Ren, Brid Ryan, John Satterlee, Andrea Wurster, and an anonymous reviewer. Thanks to all the members of the ENCODE Consortium for their dedication and enthusiasm. Finally, thanks to my NHGRI ENCODE colleagues Elise Feingold, who launched ENCODE and leads the project, Peter Good, who has been with ENCODE from the start, and Dan Gilchrist.

REFERENCES

- Allfrey VG, Faulkner R, Mirsky AE. 1964. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc Natl Acad Sci* 51: 786–794.
- Allis CD, Jenuwein T, Reinberg D, Caparros ML. 2007. *Epigenetics* Cold Spring Harbor Laboratory Press.
- Batut P, Dobin A, Plessy C, Carninci P, Gingeras TR. 2013. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res* 23: 169–180.
- Bedell MA, Brannan CI, Evans EP, Copeland NG, Jenkins NA, Donovan PJ. 1995. DNA rearrangements located over 100 kb 5' of the Steel (Sl)-coding region in Steel-panda and Steel-contrasted mice deregulate Sl expression and cause female sterility by disrupting ovarian follicle development. *Genes Dev* 9: 455–470.
- Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK. 2011a. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol* 12: R10.
- Bell O, Tiwari VK, Thoma NH, Schubeler D. 2011b. Determinants and dynamics of genome accessibility. *Nat Rev Genet* 12: 554–564.
- Biggin MD. 2011. Animal transcription networks as highly connected, quantitative continua. *Dev Cell* 21: 611–626.
- Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799–816.
- Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M. 2007. Divergence of transcription factor binding sites across related yeast species. *Science* 317: 815–819.
- Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, et al. 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 22: 1790–1797.
- Campos EI, Reinberg D. 2009. Histones: annotating chromatin. *Annu Rev Genet* 43: 559–599.
- Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, et al. 2009. Unlocking the secrets of the genome. *Nature* 459: 927–930.
- Consortium TEP. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306: 636–640.
- Consortium TEP. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* 9: e1001046.
- Consortium TEP. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74.
- Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci* 107: 21931–21936.
- Davison LJ, Wallace C, Cooper JD, Cope NF, Wilson NK, Smyth DJ, Howson JM, Saleh N, Al-Jeffery A, Angus KL, et al. 2012. Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene. *Hum Mol Genet* 21: 322–333.
- Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, et al. 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482: 390–394.
- Diaz P, Cado D, Winoto A. 1994. A locus control region in the T cell receptor α/δ locus. *Immunity* 1: 207–217.
- Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* 489: 101–108.
- Edwards SL, Beesley J, French JD, Dunning AM. 2013. Beyond GWAS: illuminating the dark road from association to function. *Am J Hum Genet* 93: 779–797.
- Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* 28: 817–825.
- Ernst J, Kheradpour P, Mikkelsen TS, Shores N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473: 43–49.
- Festenstein R, Tolaini M, Corbella P, Mamalaki C, Parrington J, Fox M, Miliou A, Jones M, Kioussis D. 1996. Locus control region function and heterochromatin-induced position effect variegation. *Science* 271: 1123–1125.
- Fisher WW, Li JJ, Hammonds AS, Brown JB, Pfeiffer BD, Weiszmam R, MacArthur S, Thomas S, Stamatoyannopoulos JA, Eisen MB, et al. 2012. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proc Natl Acad Sci* 109: 21330–21335.
- Forrester WC, Takegawa S, Papayannopoulou T, Stamatoyannopoulos G, Groudine M. 1987. Evidence for a locus activation region: the formation of developmentally stable hypersensitive sites in globin-expressing hybrids. *Nucleic Acids Res* 15: 10159–10177.
- Frank J, Pignata C, Panteleyev AA, Prowse DM, Baden H, Weiner L, Gaetaniello L, Ahmad W, Pozzi N, Cserhalmi-Friedman PB, et al. 1999. Exposing the human nude phenotype. *Nature* 398: 473–474.
- Fraser HB. 2013. Gene expression drives local adaptation in humans. *Gen Res* 23: 1089–1096.
- Fraser P, Pruzina S, Antoniou M, Grosveld F. 1993. Each hypersensitive site of the human β -globin locus control region confers a different developmental pattern of expression on the globin genes. *Genes Dev* 7: 106–113.
- Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, et al. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature* 489: 91–100.
- Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* 330: 1775–1787.
- Gilmour DS, Lis JT. 1984. Detecting protein–DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proc Natl Acad Sci* 81: 4275–4279.
- Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471: 473–479.
- Gross DS, Garrard WT. 1988. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* 57: 159–197.
- Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, et al. 2013. Identifying recent adaptations in large-scale genomic data. *Cell* 152: 703–713.

- Hardison RC. 2012. Genome-wide epigenetic data facilitate understanding of disease susceptibility association studies. *J Biol Chem* 287: 30932–30940.
- Hardison RC, Taylor J. 2012. Genomic approaches towards finding *cis*-regulatory modules in animals. *Nat Rev Genet* 13: 469–483.
- Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* 22: 1760–1774.
- Hebbes TR, Clayton AL, Thorne AW, Crane-Robinson C. 1994. Core histone hyperacetylation co-maps with generalized DNase I sensitivity in the chicken β -globin chromosomal domain. *EMBO J* 13: 1823–1830.
- Hebbes TR, Thorne AW, Crane-Robinson C. 1988. A direct link between core histone acetylation and transcriptionally active chromatin. *EMBO J* 7: 1395–1402.
- Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459: 108–112.
- Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 39: 311–318.
- Henikoff S, Shilatifard A. 2011. Histone modification: cause or cog? *Trends Genet* 27: 389–396.
- Herdewyn S, Zhao H, Moisse M, Race V, Matthijs G, Reumers J, Kusters B, Schelhaas HJ, van den Berg LH, Goris A, et al. 2012. Whole-genome sequencing reveals a coding non-pathogenic variant tagging a non-coding pathogenic hexanucleotide repeat expansion in C9orf72 as cause of amyotrophic lateral sclerosis. *Hum Mol Genet* 21: 2412–2419.
- Hu X, Kim H, Stahl E, Plenge R, Daly M, Raychaudhuri S. 2011. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *Am J Hum Genet* 89: 496–506.
- Huff JT, Plocik AM, Guthrie C, Yamamoto KR. 2010. Reciprocal intronic and exonic histone modification regions in humans. *Nat Struct Mol Biol* 17: 1495–1499.
- Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409: 533–538.
- Johnson KD, Bresnick EH. 2002. Dissecting long-range transcriptional mechanisms by chromatin immunoprecipitation. *Methods* 26: 27–36.
- Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484: 55–61.
- Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316: 1484–1488.
- Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, et al. 2011. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* 471: 480–485.
- Kheradpour P, Kellis M. 2013. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res*.
- Kleinjan DA, Lettice LA. 2008. Long-range gene control and genetic disease. *Adv Genet* 61: 339–388.
- Lee GR, Fields PE, Griffin TJ, Flavell RA. 2003. Regulation of the Th2 cytokine locus by a locus control region. *Immunity* 19: 145–153.
- Lettice LA, Horikoshi T, Heaney SJ, van Baren MJ, van der Linde HC, Breedveld GJ, Joosse M, Akarsu N, Oostra BA, Endo N, et al. 2002. Disruption of a long-range *cis*-acting regulator for Shh causes preaxial polydactyly. *Proc Natl Acad Sci* 99: 7548–7553.
- Levinger L, Varshavsky A. 1980. High-resolution fractionation of nucleosomes: minor particles, “whiskers,” and separation of mononucleosomes containing and lacking A24 semihistone. *Proc Natl Acad Sci* 77: 3244–3248.
- Li CC, Ramirez-Carrozzi VR, Smale ST. 2006. Pursuing gene regulation ‘logic’ via RNA interference and chromatin immunoprecipitation. *Nat Immunol* 7: 692–697.
- Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, et al. 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148: 84–98.
- Lin R, Leone JW, Cook RG, Allis CD. 1989. Antibodies specific to acetylated histones document the existence of deposition- and transcription-related histone acetylation in Tetrahymena. *J Cell Biol* 108: 1577–1588.
- Litt MD, Simpson M, Gaszner M, Allis CD, Felsenfeld G. 2001. Correlation between histone lysine methylation and developmental changes at the chicken β -globin locus. *Science* 293: 2453–2455.
- Livvyatan I, Harikumar A, Nissim-Rafinia M, Duttagupta R, Gingeras TR, Meshorer E. 2013. Non-polyadenylated transcription in embryonic stem cells reveals novel non-coding RNA related to pluripotency and differentiation. *Nucleic Acids Res* 41: 6300–6315.
- Loots GG, Kneissel M, Keller H, Baptist M, Chang J, Collette NM, Ovcharenko D, Plajzer-Frick I, Rubin EM. 2005. Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease. *Genome Res* 15: 928–935.
- MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335: 823–828.
- Maston GA, Landt SG, Snyder M, Green MR. 2012. Characterization of enhancer function from genome-wide analyses. *Annu Rev Genomics Hum Genet* 13: 29–57.
- Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337: 1190–1195.
- Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, et al. 2013. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* 41: D64–D69.
- Mohrs M, Blankespoor CM, Wang ZE, Loots GG, Afzal V, Hadeiba H, Shinkai K, Rubin EM, Locksley RM. 2001. Deletion of a coordinate regulator of type 2 cytokine expression in mice. *Nat Immunol* 2: 842–847.
- Mortlock DP, Pregizer S. 2012. Identifying functional annotation for non-coding genomic sequences. *Curr Protoc Hum Genet*.
- Nardone J, Lee DU, Ansel KM, Rao A. 2004. Bioinformatics for the ‘bench biologist’: how to find regulatory regions in genomic DNA. *Nat Immunol* 5: 768–774.
- Negre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, et al. 2011. A *cis*-regulatory map of the *Drosophila* genome. *Nature* 471: 527–531.
- Nehls M, Kyewski B, Messerle M, Waldschutz R, Schuddekopf K, Smith AJ, Boehm T. 1996. Two genetically separable steps in the differentiation of thymic epithelium. *Science* 272: 886–889.
- Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA. 2012a. Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 150: 1274–1286.
- Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, et al. 2012b. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489: 83–90.
- Ng JW, Barrett LM, Wong A, Kuh D, Smith GD, Relton CL. 2012. The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities. *Genome Biol* 13: 246.
- Noonan JP, McCallion AS. 2010. Genomics of long-range regulatory elements. *Annu Rev Genomics Hum Genet* 11: 1–23.
- Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* 39: 730–732.
- Ozsolak F, Poling LL, Wang Z, Liu H, Liu XS, Roeder RG, Zhang X, Song JS, Fisher DE. 2008. Chromatin structure analyses identify miRNA promoters. *Genes Dev* 22: 3172–3183.
- Pazin MJ, Kamakaka RT, Kadonaga JT. 1994. ATP-dependent nucleosome reconfiguration and transcriptional activation from preassembled chromatin templates. *Science* 266: 2007–2011.
- Pazin MJ, Sheridan PL, Cannon K, Cao Z, Keck JG, Kadonaga JT, Jones KA. 1996. NF- κ B-mediated chromatin reconfiguration and transcriptional activation of the HIV-1 enhancer in vitro. *Genes Dev* 10: 37–49.

- Penagarikano O, Mulle JG, Warren ST. 2007. The pathophysiology of Fragile X syndrome. *Annu Rev Genomics Hum Genet* 8: 109–129.
- Pickrell JK. 2013. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *bioRxiv*.
- Ponting CP, Hardison RC. 2011. What fraction of the human genome is functional? *Genome Res* 21: 1769–1776.
- Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470: 279–283.
- Rakyan VK, Down TA, Balding DJ, Beck S. 2011. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 12: 529–541.
- Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* 290: 2306–2309.
- Rivera CM, Ren B. 2013. Mapping human epigenomes. *Cell* 155: 39–55.
- Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 330: 1787–1797.
- Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. *Nature* 489: 109–113.
- Sawada S, Scarborough JD, Killeen N, Littman DR. 1994. A lineage-specific transcriptional silencer regulates CD4 gene expression during T lymphocyte development. *Cell* 77: 917–929.
- Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. 2012. Linking disease associations with regulatory information in the human genome. *Genome Res* 22: 1748–1759.
- Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 328: 1036–1040.
- Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, Crawford GE, Furey TS. 2013. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res* 23: 777–788.
- Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, et al. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci* 100: 15776–15781.
- Siu G, Wurster AL, Duncan DD, Soliman TM, Hedrick SM. 1994. A transcriptional silencer controls the developmental expression of the CD4 gene. *EMBO J* 13: 3570–3579.
- Solomon MJ, Larsen PL, Varshavsky A. 1988. Mapping protein–DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* 53: 937–947.
- Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, et al. 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 480: 490–495.
- Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T, et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* 13: 418.
- Stergachis AB, Neph S, Reynolds A, Humbert R, Miller B, Paige SL, Vernot B, Cheng JB, Thurman RE, Sandstrom R, et al. 2013. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* 154: 888–903.
- Suganuma T, Workman JL. 2011. Signals and combinatorial functions of histone modifications. *Annu Rev Biochem* 80: 473–499.
- Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* 489: 75–82.
- Tilgner H, Knowles DG, Johnson R, Davis CA, Chakraborty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigo R. 2012. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* 22: 1616–1625.
- Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* 39: 31–40.
- Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, Raychaudhuri S. 2013. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* 45: 124–130.
- Turner BM. 1989. Acetylation and deacetylation of histone H4 continue through metaphase with depletion of more-acetylated isoforms and altered site usage. *Exp Cell Res* 182: 206–214.
- Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, Cross MK, Williams BA, Stamatoyannopoulos JA, Crawford GE, et al. 2013. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res* 23: 555–567.
- Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457: 854–858.
- Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, et al. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* 22: 1798–1812.
- Ward LD, Kellis M. 2012a. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337: 1675–1678.
- Ward LD, Kellis M. 2012b. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 40: D930–D934.
- Ward LD, Kellis M. 2012c. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* 30: 1095–1106.
- Weinmann AS, Farnham PJ. 2002. Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. *Methods* 26: 37–47.
- Weintraub H, Groudine M. 1976. Chromosomal subunits in active genes have an altered conformation. *Science* 193: 848–856.
- Wu C. 1980. The 5' ends of *Drosophila* heat shock genes in chromatin are hypersensitive to DNase I. *Nature* 286: 854–860.
- Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M, et al. 2012. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* 13: R48.

Cold Spring Harbor Protocols

Using the ENCODE Resource for Functional Annotation of Genetic Variants

Michael J. Pazin

Cold Spring Harb Protoc; doi: 10.1101/pdb.top084988 originally published online March 11, 2015

Email Alerting Service

Receive free email alerts when new articles cite this article - [click here](#).

Subject Categories

Browse articles on similar topics from *Cold Spring Harbor Protocols*.

[Bioinformatics/Genomics, general](#) (192 articles)
[Computational Biology](#) (100 articles)
[Gene Prediction](#) (12 articles)
[Genetic Variation](#) (86 articles)
[Genetics, general](#) (374 articles)
[Genome Analysis](#) (191 articles)

To subscribe to *Cold Spring Harbor Protocols* go to:
<http://cshprotocols.cshlp.org/subscriptions>