A User's Guide to the Encyclopedia of DNA Elements (ENCODE)

The ENCODE Project Consortium[¶]*

Abstract

The mission of the Encyclopedia of DNA Elements (ENCODE) Project is to enable the scientific and medical communities to interpret the human genome sequence and apply it to understand human biology and improve health. The ENCODE Consortium is integrating multiple technologies and approaches in a collective effort to discover and define the functional elements encoded in the human genome, including genes, transcripts, and transcriptional regulatory regions, together with their attendant chromatin states and DNA methylation patterns. In the process, standards to ensure high-quality data have been implemented, and novel algorithms have been developed to facilitate analysis. Data and derived results are made available through a freely accessible database. Here we provide an overview of the project and the resources it is generating and illustrate the application of ENCODE data to interpret the human genome.

Citation: The ENCODE Project Consortium (2011) A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol 9(4): e1001046. doi:10.1371/journal.pbio.1001046

Academic Editor: Peter B. Becker, Adolf Butenandt Institute, Germany

Received September 23, 2010; **Accepted** March 10, 2011; **Published** April 19, 2011

Funding: Funded by the National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. The role of the NIH Project Management Group in the preparation of this paper was limited to coordination and scientific management of the ENCODE Consortium.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: 3C, Chromosome Conformation Capture; API, application programming interface; CAGE, Cap-Analysis of Gene Expression; ChIP, chromatin immunoprecipitation; DCC, Data Coordination Center; DHS, DNaseI hypersensitive site; ENCODE, Encyclopedia of DNA Elements; EPO, Enredo, Pecan, Ortheus approach; FDR, false discovery rate; GEO, Gene Expression Omnibus; GWAS, genome-wide association studies; IDR, Irreproducible Discovery Rate; Methyl-seq, sequencing-based methylation determination assay; NHGRI, National Human Genome Research Institute; PASRs, promoter-associated short RNAs; PET, Paired-End diTag; RACE, Rapid Amplification of cDNA Ends; RNA Pol2, RNA polymerase 2; RBP, RNA-binding protein; RRBS, Reduced Representation Bisulfite Sequencing; SRA, Sequence Read Archive; TAS, trait/disease-associated SNP; TF, transcription factor; TSS, transcription start site

* E-mail: rmyers@hudsonalpha.org (RMM); jstam@u.washington.edu (JS); mpsnyder@stanford.edu (MS); dunham@ebi.ac.uk (ID); rch8@psu.edu (RCH); bernstein.bradley@mgh.harvard.edu (BEB); gingeras@cshl.edu (TRG); kent@soe.ucsc.edu (WJK); birney@ebi.ac.uk (EB); woldb@caltech.edu (BW); greg.crawford@duke.edu (GEC)

¶ Membership of the ENCODE Project Consortium is provided in the Acknowledgments.

I. Introduction and Project Overview

Interpreting the human genome sequence is one of the leading challenges of 21st century biology [1]. In 2003, the National Human Genome Research Institute (NHGRI) embarked on an ambitious project—the Encyclopedia of DNA Elements (ENCODE)—aiming to delineate all of the functional elements encoded in the human genome sequence [2]. To further this goal, NHGRI organized the ENCODE Consortium, an international group of investigators with diverse backgrounds and expertise in production and analysis of high-throughput functional genomic data. In a pilot project phase spanning 2003–2007, the Consortium applied and compared a variety of experimental and computational methods to annotate functional elements in a defined 1% of the human genome [3]. Two additional goals of the pilot ENCODE Project were to develop and advance technologies for annotating the human genome, with the combined aims of achieving higher accuracy, completeness, and cost-effective throughput and establishing a paradigm for sharing functional genomics data. In 2007, the ENCODE Project was expanded to study the entire human genome, capitalizing on experimental and computational technology developments during the pilot project period. Here we describe this expanded project, which we refer to throughout as the ENCODE Project, or ENCODE.

The major goal of ENCODE is to provide the scientific community with high-quality, comprehensive annotations of candidate functional elements in the human genome. For the purposes of this article, the term "functional element" is used to denote a discrete region of the genome that encodes a defined product (e.g., protein) or a reproducible biochemical signature, such as transcription or a specific chromatin structure. It is now widely appreciated that such signatures, either alone or in combinations, mark genomic sequences with important functions, including exons, sites of RNA processing, and transcriptional regulatory elements such as promoters, enhancers, silencers, and insulators. However, it is also important to recognize that while certain biochemical signatures may be associated with specific functions, our present state of knowledge may not yet permit definitive declaration of the ultimate biological role(s), function(s), or mechanism(s) of action of any given genomic element.

At present, the proportion of the human genome that encodes functional elements is unknown. Estimates based on comparative genomic analyses suggest that 3%–8% of the base pairs in the human genome are under purifying (or negative) selection [4–7]. However, this likely underestimates the prevalence of functional features, as current comparative methods may not account for lineage-specific evolutionary innovations, functional elements that are very small or fragmented [8], elements that are rapidly evolving or subject to nearly neutral evolutionary processes, or elements that lie in repetitive regions of the genome.

The current phase of the ENCODE Project has focused on completing two major classes of annotations: genes (both protein-coding and non-coding) and their RNA transcripts, and transcriptional regulatory regions. To accomplish these

Author Summary

The Encyclopedia of DNA Elements (ENCODE) Project was created to enable the scientific and medical communities to interpret the human genome sequence and to use it to understand human biology and improve health. The ENCODE Consortium, a large group of scientists from around the world, uses a variety of experimental methods to identify and describe the regions of the 3 billion base-pair human genome that are important for function. Using experimental, computational, and statistical analyses, we aimed to discover and describe genes, transcripts, and transcriptional regulatory regions, as well as DNA binding proteins that interact with regulatory regions in the genome, including transcription factors, different versions of histones and other markers, and DNA methylation patterns that define states of the genome in various cell types. The ENCODE Project has developed standards for each experiment type to ensure high-quality, reproducible data and novel algorithms to facilitate analysis. All data and derived results are made available through a freely accessible database. This article provides an overview of the complete project and the resources it is generating, as well as examples to illustrate the application of ENCODE data as a user's guide to facilitate the interpretation of the human genome.

goals, seven ENCODE Data Production Centers encompassing 27 institutions have been organized to focus on generating multiple complementary types of genome-wide data (Figure 1 and Figure S1). These data include identification and quantification of RNA species in whole cells and in sub-cellular compartments, mapping of protein-coding regions, delineation of chromatin and DNA accessibility and structure with nucleases and chemical probes, mapping of histone modifications and transcription factor (TF) binding sites by chromatin immunoprecipitation (ChIP), and measurement of DNA methylation (Figure 2 and Table 1). In parallel with the major production efforts, several smaller-scale efforts are examining long-range chromatin interactions, localizing binding proteins on RNA, identifying transcriptional silencer elements, and understanding detailed promoter sequence architecture in a subset of the genome (Figure 1 and Table 1).

ENCODE has placed emphasis on data quality, including ongoing development and application of standards for data reproducibility and the collection of associated experimental information (i.e., metadata). Adoption of state-of-the-art, massively parallel DNA sequence analysis technologies has greatly facilitated standardized data processing, comparison, and integration [9,10]. Primary and processed data, as well as relevant experimental methods and parameters, are collected by a central Data Coordination Center (DCC) for curation, quality review, visualization, and dissemination (Figure 1). The Consortium releases data rapidly to the public through a web-accessible database (http://genome.ucsc.edu/ENCODE/) [11] and provides a visualization framework and analytical tools to facilitate use of the data [12], which are organized into a web portal (http://encodeproject.org).

To facilitate comparison and integration of data, ENCODE data production efforts have prioritized selected sets of cell types (Table 2). The highest priority set (designated "Tier 1") includes two widely studied immortalized cell lines—K562 erythroleukemia cells [13]; an EBV-immortalized B-lymphoblastoid line (GM12878, also being studied by the 1,000 Genomes Project; http://1000genomes.org) and the H1 human embryonic stem cell line [14]. A secondary priority set (Tier 2) includes HeLa-S3 cervical carcinoma cells [15], HepG2 hepatoblastoma cells [16], and primary (non-transformed) human umbilical vein endothelial cells (HUVEC; [17]), which have limited proliferation potential in culture. To capture a broader spectrum of human biological diversity, a third set (Tier 3) currently comprises more than 100 cell types that are being analyzed in selected assays (Table 2). Standardized growth conditions for all ENCODE cell types have been established and are available through the ENCODE web portal (http://encodeproject.org, "cell types" link).

This report is intended to provide a guide to the data and resources generated by the ENCODE Project to date on Tier 1–3 cell types. We summarize the current state of ENCODE by describing the experimental and computational approaches used to generate and analyze data. In addition, we outline how to access datasets and provide examples of their use.

II. ENCODE Project Data

The following sections describe the different types of data being produced by the ENCODE Project (Table 1).

Genes and Transcripts

Gene annotation. A major goal of ENCODE is to annotate all protein-coding genes, pseudogenes, and non-coding transcribed loci in the human genome and to catalog the products of transcription including splice isoforms. Although the human genome contains ~20,000 protein-coding genes [18], accurate identification of all protein-coding transcripts has not been straightforward. Annotation of pseudogenes and noncoding transcripts also remains a considerable challenge. While automatic gene annotation algorithms have been developed, manual curation remains the approach that delivers the highest level of accuracy, completeness, and stability [19]. The ENCODE Consortium has therefore primarily relied on manual curation with moderate implementation of automated algorithms to produce gene and transcript models that can be verified by traditional experimental and analytical methods. This annotation process involves consolidation of all evidence of transcripts (cDNA, EST sequences) and proteins from public databases, followed by building gene structures based on supporting experimental data [20]. More than 50% of annotated transcripts have no predicted coding potential and are classified by ENCODE into different transcript categories. A classification that summarizes the certainty and types of the annotated structures is provided for each transcript (see http://www.gencodegenes.org/biotypes.html for details). The annotation also includes extensive experimental validation by RT-PCR for novel transcribed loci (i.e., those not previously observed and deposited into public curated databases such as RefSeq). Pseudogenes are identified primarily by a combination of similarity to other protein-coding genes and an obvious functional disablement such as an in-frame stop codon. Because it is difficult to validate pseudogenes experimentally, three independent annotation methods from Yale ("pseudopipe") [21], UCSC ("retrofinder"; http://users.soe.ucsc.edu/~markd/gene-sets-new/pseudoGenes/RetroFinder.html, and references therein), and the Sanger Center [20] are combined to produce a consensus pseudogene set. Ultimately, each gene or transcript model is assigned one of three confidence levels. Level 1 includes genes validated by RT-PCR and sequencing, plus consensus pseudogenes. Level 2 includes manually annotated coding and long non-coding loci that have transcriptional evidence in EMBL/GenBank. Level 3 includes Ensembl gene predictions in regions not yet manually annotated or for which there is new transcriptional evidence.

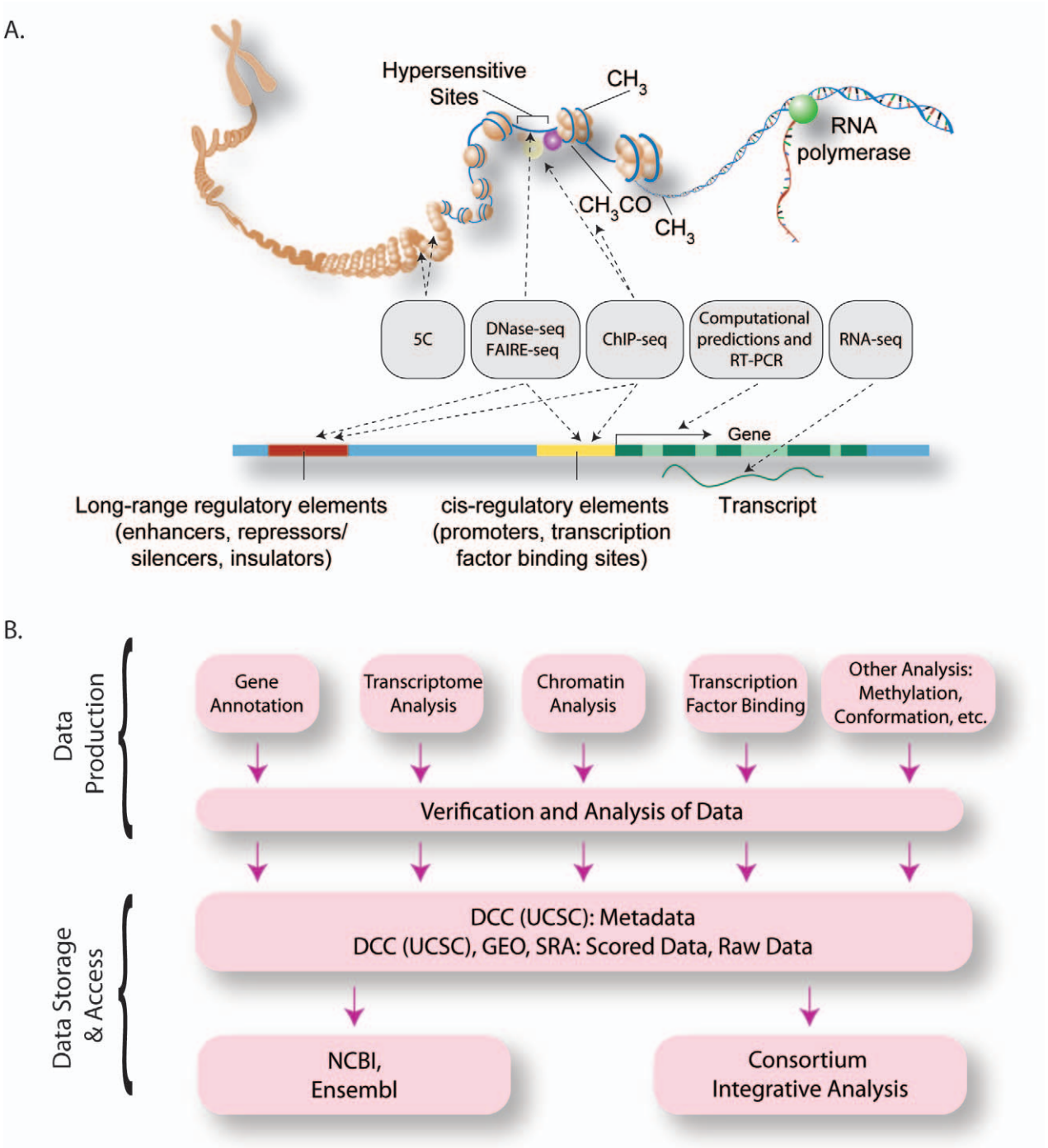

Figure 1. The Organization of the ENCODE Consortium. (A) Schematic representation of the major methods that are being used to detect functional elements (gray boxes), represented on an idealized model of mammalian chromatin and a mammalian gene. (B) The overall data flow from the production groups after reproducibility assessment to the Data Coordinating Center (UCSC) for public access and to other public databases. Data analysis is performed by production groups for quality control and research, as well as at a cross-Consortium level for data integration.
doi:10.1371/journal.pbio.1001046.g001

The result of ENCODE gene annotation (termed "GENCODE") is a comprehensive catalog of transcripts and gene models. ENCODE gene and transcript annotations are updated bimonthly and are available through the UCSC ENCODE browser, distributed annotation servers (DAS; see http://genome.ucsc.edu/cgi-bin/das/hg18/features?segment=21:33031597,33041570;type=wgEncodeGencodeManualV3), and the Ensembl Browser [22].

RNA transcripts. ENCODE aims to produce a comprehensive genome-wide catalog of transcribed loci that characterizes the size, polyadenylation status, and subcellular compartmentalization of all transcripts (Table 1).

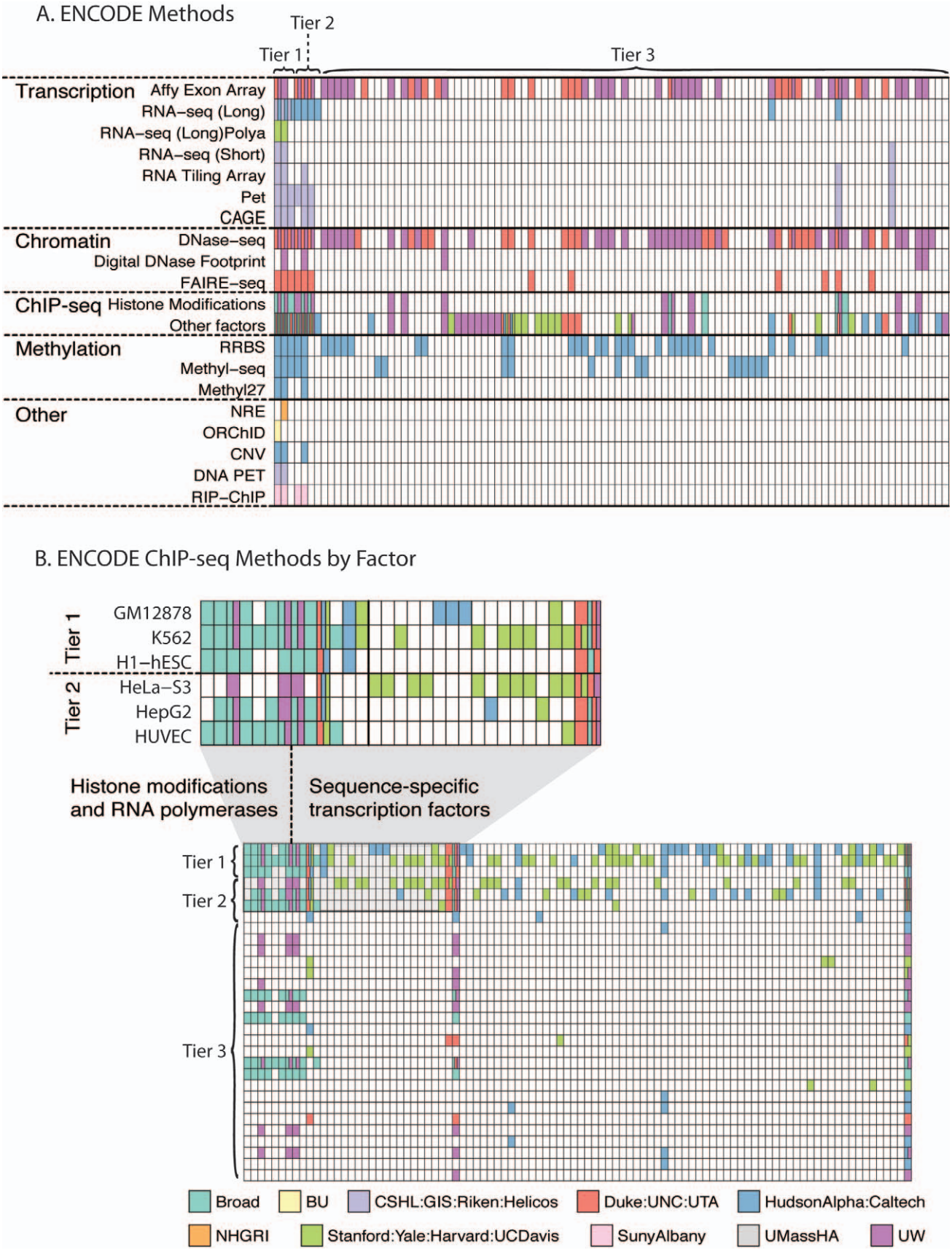
A. ENCODE Methods
Tier 1
Tier 2
Tier 3
Transcription
Affy Exon Array
RNA–seq (Long)
RNA–seq (Long)Polya
RNA–seq (Short)
RNA Tiling Array
Pet
CAGE
Chromatin
DNase–seq
Digital DNase Footprint
FAIRE–seq
ChIP-seq
Histone Modifications
Other factors
Methylation
RRBS
Methyl–seq
Methyl27
Other
NRE
ORChID
CNV
DNA PET
RIP–ChIP
B. ENCODE ChIP-seq Methods by Factor
Tier 1
GM12878
K562
H1–hESC
Tier 2
HeLa–S3
HepG2
HUVEC
Histone modifications and RNA polymerases
Sequence-specific transcription factors
Tier 1
Tier 2
Tier 3
Broad
BU
CSHL:GIS:Riken:Helicos
Duke:UNC:UTA
HudsonAlpha:Caltech
NHGRI
Stanford:Yale:Harvard:UCDavis
SunyAlbany
UMassHA
UW

Figure 2. Data available from the ENCODE Consortium. (A) A data matrix representing all ENCODE data types. Each row is a method and each column is a cell line on which the method could be applied to generate data. Colored cells indicate that data have been generated for that method on that cell line. The different colors represent data generated from different groups in the Consortium as indicated by the key at the bottom of the figure. In some cases, more than one group has generated equivalent data; these cases are indicated by subdivision of the cell to accommodate multiple colors. (B) Data generated by ChIP-seq are split into a second matrix where the cells now represent cell types (rows) split by the factor or histone modification to which the antibody is raised (columns). The colors again represent the groups as indicated by the key. The upper left corner of this matrix has been expanded immediately above the panel to better illustrate the data. All data were collected from the ENCODE public download repository at http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC on September 1, 2010.
doi:10.1371/journal.pbio.1001046.g002

ENCODE has generated transcript data with high-density (5 bp) tiling DNA microarrays [23] and massively parallel DNA sequencing methods [9,10,24], with the latter predominating in ongoing efforts. Both polyA+ and polyA− RNAs are being analyzed. Because subcellular compartmentalization of RNAs is important in RNA processing and function, such as nuclear retention of unspliced coding transcripts [25] or snoRNA activity in the nucleolus [26], ENCODE is analyzing not only total whole cell RNAs but also those concentrated in the nucleus and cytosol. Long (>200 nt) and short RNAs (<200 nt) are being sequenced from each subcellular compartment, providing catalogs of potential miRNAs, snoRNA, promoter-associated short RNAs (PASRs) [27], and other short cellular RNAs. Total RNA from K562 and GM12878 cells has been mapped by hybridization to high-density tiling arrays and sequenced to a depth of >500 million paired-end 76 bp reads under conditions where the strand of the RNA transcript is determined, providing considerable depth of transcript coverage (see below).

These analyses reveal that the human genome encodes a diverse array of transcripts. For example, in the proto-oncogene *TP53* locus, RNA-seq data indicate that, while *TP53* transcripts are accurately assigned to the minus strand, those for the oppositely transcribed, adjacent gene *WRAP53* emanate from the plus strand (Figure 3). An independent transcript within the first intron of *TP53* is also observed in both GM12878 and K562 cells (Figure 3).

Additional transcript annotations include exonic regions and splice junctions, transcription start sites (TSSs), transcript 3′ ends, spliced RNA length, locations of polyadenylation sites, and locations with direct evidence of protein expression. TSSs and 3′ ends of transcripts are being determined with two approaches, Paired-End diTag (PET) [28] and Cap-Analysis of Gene Expression (CAGE) [29–31] sequencing.

Table 1. Experimental assays used by the ENCODE Consortium.

Gene/Transcript Analysis		
Region/Feature	**Method**	**Group**
Gene annotation	GENCODE	Wellcome Trust
PolyA+ coding regions	RNA-seq; tiling DNA microarrays; PET	CSHL; Stanford/Yale//Harvard; Caltech
Total RNA coding regions	RNA-seq; tiling DNA microarrays; PET	CSHL
Coding regions in subcellular RNA fractions (e.g. nuclear, cytoplasmic)	PET	CSHL
Small RNAs	short RNA-seq	CSHL
Transcription initiation (5′-end) and termination (3-end′) sites	CAGE; diTAGs	RIKEN, GIS
Full-length RNAs	RACE	University of Geneva; University of Lausanne
Protein-bound RNA coding regions	RIP; CLIP	SUNY-Albany; CSHL
Transcription Factors/Chromatin		
Elements/Regions	**Method(s)**	**Group(s)**
Transcription Factor Binding Sites (TFBS)	ChIP-seq	Stanford/Yale/UC-Davis/Harvard; HudsonAlpha/Caltech; Duke/UT-Austin; UW; U. Chicago/Stanford
Chromatin structure (accessibility, etc.)	DNaseI hypersensitivity; FAIRE	UW; Duke; UNC
Chromatin modifications (H3K27ac, H3K27me3, H3K36me3, etc.)	ChIP-seq	Broad; UW
DNaseI footprints	Digital genomic footprinting	UW
Other Elements/Features		
Feature	**Method(s)**	**Group(s)**
DNA methylation	RRBS; Illumina Methyl27; Methyl-seq	HudsonAlpha
Chromatin interactions	5C; CHIA-PET	UMass; UW; GIS
Genotyping	Illumina 1M Duo	HudsonAlpha

doi:10.1371/journal.pbio.1001046.t001

Table 2. ENCODE cell types.

Cell Type	Tier	Description	Source
GM12878	1	B-Lymphoblastoid cell line	Coriell GM12878
K562	1	Chronic Myelogenous/Erythroleukemia cell line	ATCC CCL-243
H1-hESC	1	Human Embryonic Stem Cells, line H1	Cellular Dynamics International
HepG2	2	Hepatoblastoma cell line	ATCC HB-8065
HeLa-S3	2	Cervical carcinoma cell line	ATCC CCL-2.2
HUVEC	2	Human Umbilical Vein Endothelial Cells	Lonza CC-2517
Various (Tier 3)	3	Various cell lines, cultured primary cells, and primary tissues	Various

doi:10.1371/journal.pbio.1001046.t002

Transcript annotations throughout the genome are further corroborated by comparing tiling array data with deep sequencing data and by the manual curation described above. Additionally, selected compartment-specific RNA transcripts that cannot be mapped to the current build of the human genome sequence have been evaluated by 5′/3′ Rapid Amplification of cDNA Ends (RACE) [32], followed by RT-PCR cloning and sequencing. To assess putative protein products generated from novel RNA transcripts and isoforms, proteins may be sequenced and quantified by mass spectrometry and mapped back to their encoding transcripts [33,34]. ENCODE has recently begun to study proteins from distinct subcellular compartments of K562 and GM12878 cells by using this complementary approach.

Cis-Regulatory Regions

Cis-regulatory regions include diverse functional elements (e.g., promoters, enhancers, silencers, and insulators) that collectively modulate the magnitude, timing, and cell-specificity of gene expression [35]. The ENCODE Project is using multiple approaches to identify *cis*-regulatory regions, including localizing their characteristic chromatin signatures and identifying sites of occupancy of sequence-specific transcription factors. These approaches are being combined to create a comprehensive map of human *cis*-regulatory regions.

Chromatin structure and modification. Human *cis*-regulatory regions characteristically exhibit nuclease hypersensitivity [36–39] and may show increased solubility after chromatin fixation and fragmentation [40,41]. Additionally, specific patterns of post-translational histone modifications [42,43] have been connected with distinct classes of regions such as promoters and enhancers [3,44–47] as well as regions subject to programmed repression by Polycomb complexes [48,49] or other mechanisms [46,50,51]. Chromatin accessibility and histone modifications thus provide independent and complementary annotations of human regulatory DNA, and massively parallel, high-throughput DNA sequencing methods are being used by ENCODE to map these features on a genome-wide scale (Figure 2 and Table 1).

DNaseI hypersensitive sites (DHSs) are being mapped by two techniques: (i) capture of free DNA ends at in vivo DNaseI cleavage sites with biotinylated adapters, followed by digestion with a TypeIIS restriction enzyme to generate ~20 bp DNaseI

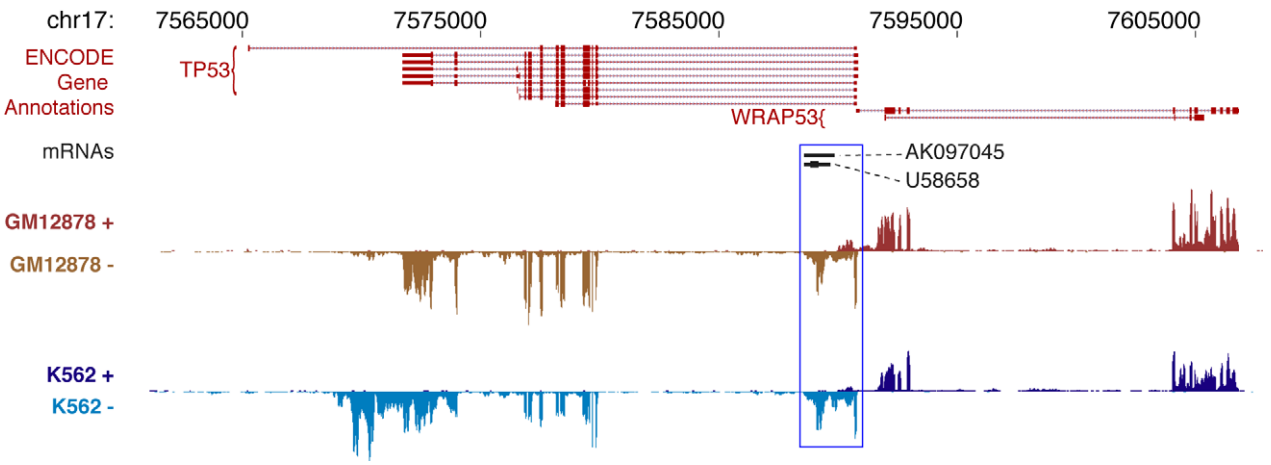

Figure 3. ENCODE gene and transcript annotations. The image shows selected ENCODE and other gene and transcript annotations in the region of the human *TP53* gene (region chr17:7,560,001–7,610,000 from the Human February 2009 (GRCh37/hg19) genome assembly). The annotated isoforms of *TP53* RNAs listed from the ENCODE Gene Annotations (GENCODE) are shown in the top tracks of the figure, along with annotation of the neighboring *WRAP53* gene. In black are two mRNA transcripts (U58658/AK097045) from GenBank. The bottom two tracks show the structure of the *TP53* region transcripts detected in nuclear polyadenylated poly A+ RNAs isolated from GM12878 and K562 cells. The RNA is characterized by RNA-seq and the RNAs detected are displayed according to the strand of origin (i.e. + and −). Signals are scaled and are present at each of the detected p53 exons. Signals are also evident at the U58658 [120] and AK097045 [121] regions located in the first 10 kb intron of the p53 gene (D17S2179E). The U58658/AK097045 transcripts are reported to be induced during differentiation of myeloid leukemia cells but are seen in both GM12878 and K562 cell lines. Finally the p53 isoform observed in K562 cells has a longer 3′UTR region than the isoform seen in the GM12878 cell line.
doi:10.1371/journal.pbio.1001046.g003

cleavage site tags [52,53] and (ii) direct sequencing of DNaseI cleavage sites at the ends of small (<300 bp) DNA fragments released by limiting treatment with DNaseI [54–56]. Chromatin structure is also being profiled with the FAIRE technique [40,57,58], in which chromatin from formaldehyde-crosslinked cells is sonicated in a fashion similar to ChIP and then extracted with phenol, followed by sequencing of soluble DNA fragments. An expanding panel of histone modifications (Figure 2) is being profiled by ChIP-seq [59–62]. In this method, chromatin from crosslinked cells is immunoprecipitated with antibodies to chromatin modifications (or other proteins of interest), the associated DNA is recovered, and the ends are subjected to massively parallel DNA sequencing. Control immunoprecipitations with a control IgG antibody or "input" chromatin—sonicated crosslinked chromatin that is not subjected to immune enrichment—are also sequenced for each cell type. These provide critical controls, as shearing of crosslinked chromatin may occur preferentially within certain regulatory DNA regions, typically promoters [41]. ENCODE chromatin data types are illustrated for a typical locus in Figure 4, which depicts the patterns of chromatin accessibility, DNaseI hypersensitive sites, and selected histone modifications in GM12878 cells.

For each chromatin data type, the "raw signal" is presented as the density of uniquely aligning sequence reads within 150 bp sliding windows in the human genome. In addition, some data are available as processed signal tracks in which filtering algorithms have been applied to reduce experimental noise. A variety of specialized statistical algorithms are applied to generate discrete high-confidence genomic annotations, including DHSs, broader regions of increased sensitivity to DNaseI, regions of enrichment by FAIRE, and regions with significant levels of specific histone modifications (see Tables 3 and S1). Notably, different histone modifications exhibit characteristic genomic distributions that may be either discrete (e.g., H3K4me3 over a promoter) or broad (e.g., H3K36me3 over an entire transcribed gene body). Because statistical false discovery rate (FDR) thresholds are applied to discrete annotations, the number of regions or elements identified under each assay type depends upon the threshold chosen. Optimal thresholds for an assay are typically determined by comparison to an independent and standard assay method or through reproducibility measurements (see below). Extensive validation of the detection of DNaseI hypersensitive sites is being performed independently with traditional Southern blotting, and more than 6,000 Southern images covering 224 regions in >12 cell types are available through the UCSC browser.

Transcription factor and RNA polymerase occupancy. Much of human gene regulation is determined by the binding of transcriptional regulatory proteins to their cognate sequence elements in *cis*-regulatory regions. ChIP-seq enables genome-scale mapping of transcription factor (TF) occupancy patterns in vivo [59,60,62] and is being extensively applied by ENCODE to create an atlas of regulatory factor binding in diverse cell types. ChIP-seq experiments rely on highly specific antibodies that are extensively characterized by immunoblot analysis and other criteria according

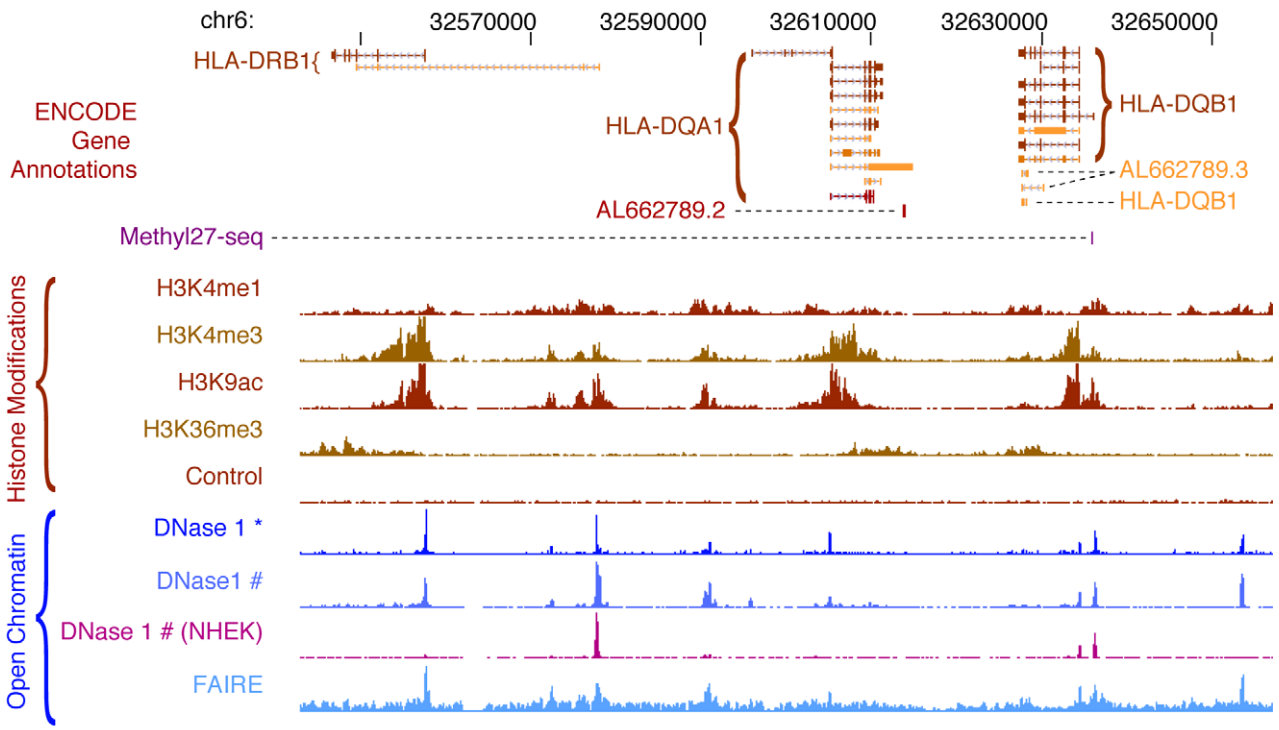

Figure 4. ENCODE chromatin annotations in the *HLA* locus. Chromatin features in a human lymphoblastoid cell line, GM12878, are displayed for a 114 kb region in the *HLA* locus. The top track shows the structures of the annotated isoforms of the *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1* genes from the ENCODE Gene Annotations (GENCODE), revealing complex patterns of alternative splicing and several non-protein-coding transcripts overlapping the protein-coding transcripts. The purple mark on the next line shows that a CpG in the promoter of the *HLA-DQB1* gene is partially methylated (assayed on the Illumina Methylation27 BeadArray platform). The densities of four histone modifications associated with transcriptionally active loci are plotted next, along with the input control signal (generated by sequencing an aliquot of the sheared chromatin for which no immunoprecipitation was performed). The last lines plot the accessibility of DNA in chromatin to nucleases (DNaseI) and reduced coverage by nucleosomes (FAIRE); peaks on these lines are DNaseI hypersensitive sites. Note that the ENCODE Consortium generates DNaseI accessibility data by two alternative protocols marked by * and #. The magenta track shows DNaseI sensitivity in a different cell line, NHEK, for comparison.
doi:10.1371/journal.pbio.1001046.g004

Table 3. Analysis tools applied by the ENCODE Consortium.

Class of Software	Description of Task	Examples[a]
Short read alignment	Computationally efficient alignment of short reads to the genome sequence	Bowtie, BWA, Maq, TopHat, GEM, STAR
Peak calling	Converting tag density to defined regions that show statistical properties consistent with binding activity	SPP, PeakSeq, Fseq, MACS, HotSpot
RNA processing	Processing RNA reads into exons and transcripts, with consideration of alternative splicing	Cufflinks, ERANGE, Fluxcapacitor
Integrative peak calling and classification	Jointly considering multiple assay signals to both define the location and character of different genomic regions	ChromHMM, Segway
Statistical tools for specific genomic tasks	Statistical methods developed for replicate-based thresholding, genome-wide-based overlap, and genome-based aggregation	IDR, GSC, ACT
Motif finding tools	Discovering the presence of sequence motifs in enriched peaks	MEME, Weeder
Data analysis frameworks	General frameworks to allow manipulation, comparison, and statistical analysis	R, Bioconductor, MatLab, Galaxy, DART, Genometools
Assign TFBS peaks to genes	Match TFBS to genes they are likely to regulate	GREAT
Compare TF binding and gene expression	Compare binding and expression; compare expressed versus nonexpressed genes	GenPattern, GSEA, Dchip
Conservation	Evaluates conservation of sequences across a range of species	phastCons, GERP, SCONE
Gene Ontology Analysis	Determine types of genes enriched for a given dataset	GO miner, BINGO, AmiGO
Network analysis	Examine relationships between genes	Cytoscape

[a]For full listings and references, see Table S1.
doi:10.1371/journal.pbio.1001046.t003

to ENCODE experimental standards. High-quality antibodies are currently available for only a fraction of human TFs, and identifying suitable immunoreagents has been a major activity of ENCODE TF mapping groups. Alternative technologies, such as epitope tagging of TFs in their native genomic context using recombineering [63,64], are also being explored.

ENCODE has applied ChIP-seq to create occupancy maps for a variety of TFs, RNA polymerase 2 (RNA Pol2) including both unphosphorylated (initiating) and phosphorylated (elongating) forms, and RNA polymerase 3 (RNA Pol3). The localization patterns of five transcription factors and RNA Pol2 in GM12878 lymphoblastoid cells are shown for a typical locus in Figure 5. Sequence reads are processed as described above for DNaseI, FAIRE, and histone modification experiments, including the application of specialized peak-calling algorithms that use input chromatin or control immunoprecipitation data to identify potential false-positives introduced by sonication or sequencing biases (Table 3). Although different peak-callers vary in performance, the strongest peaks are generally identified by multiple algorithms. Most of the sites identified by ChIP-seq are also detected by traditional ChIP-qPCR [65] or are consistent with sites reported in the literature. For example, 98% of 112 sites of CTCF occupancy previously identified by using both ChIP-chip and ChIP-qPCR [66] are also identified in ENCODE CTCF data. Whereas the binding of sequence-specific TFs is typically highly localized resulting in tight sequence tag peaks, signal from antibodies that recognize the phosphorylated (elongating) form of RNA Pol2 may detect occupancy over a wide region encompassing both the site of transcription initiation as well as the domain of elongation. Comparisons among ENCODE groups have revealed that TF and RNA Pol2 occupancy maps generated independently by different groups are highly consistent.

Additional Data Types

ENCODE is also generating additional data types to complement production projects and benchmark novel technologies. An overview of these datasets is provided in Table 1.

DNA methylation. In vertebrate genomes, methylation at position 5 of the cytosine in CpG dinucleotides is a heritable "epigenetic" mark that has been connected with both transcriptional silencing and imprinting [67,68]. ENCODE is applying several complementary approaches to measure DNA methylation. All ENCODE cell types are being assayed using two direct methods for measuring DNA methylation following sodium bisulfite conversion, which enables quantitative analysis of methylcytosines: interrogation of the methylation status of 27,000 CpGs with the Illumina Methyl27 assay [69–72] and Reduced Representation Bisulfite Sequencing (RRBS) [73], which couples *Msp*I restriction enzyme digestion, size selection, bisulfite treatment, and sequencing to interrogate the methylation status of >1,000,000 CpGs largely concentrated within promoter regions and CpG islands. Data from an indirect approach using a methylation-sensitive restriction enzyme (Methyl-seq) [74] are also available for a subset of cell types. These three approaches measure DNA methylation in defined (though overlapping) subsets of the human genome and provide quantitative determinations of the fraction of CpG methylation at each site.

DNaseI footprints. DNaseI footprinting [75] enables visualization of regulatory factor occupancy on DNA in vivo at nucleotide resolution and has been widely applied to delineate the fine structure of *cis*-regulatory regions [76]. Deep sampling of highly enriched libraries of DNaseI-released fragments (see above) enables digital quantification of per nucleotide DNaseI cleavage, which in turn enables resolution of DNaseI footprints on a large scale [55,77,78]. Digital genomic footprinting is being applied on a large scale within ENCODE to identify millions of DNaseI footprints across >12 cell types, many of which localize the specific cognate regulatory motifs for factors profiled by ChIP-seq.

Sequence and structural variation. Genotypic and structural variations within all ENCODE cell types are being interrogated at ~1 million positions distributed approximately every 1.5 kb along the human genome, providing a finely grained map of allelic variation and sequence copy number gains and losses. Genotyping data are generated with the Illumina Infinium

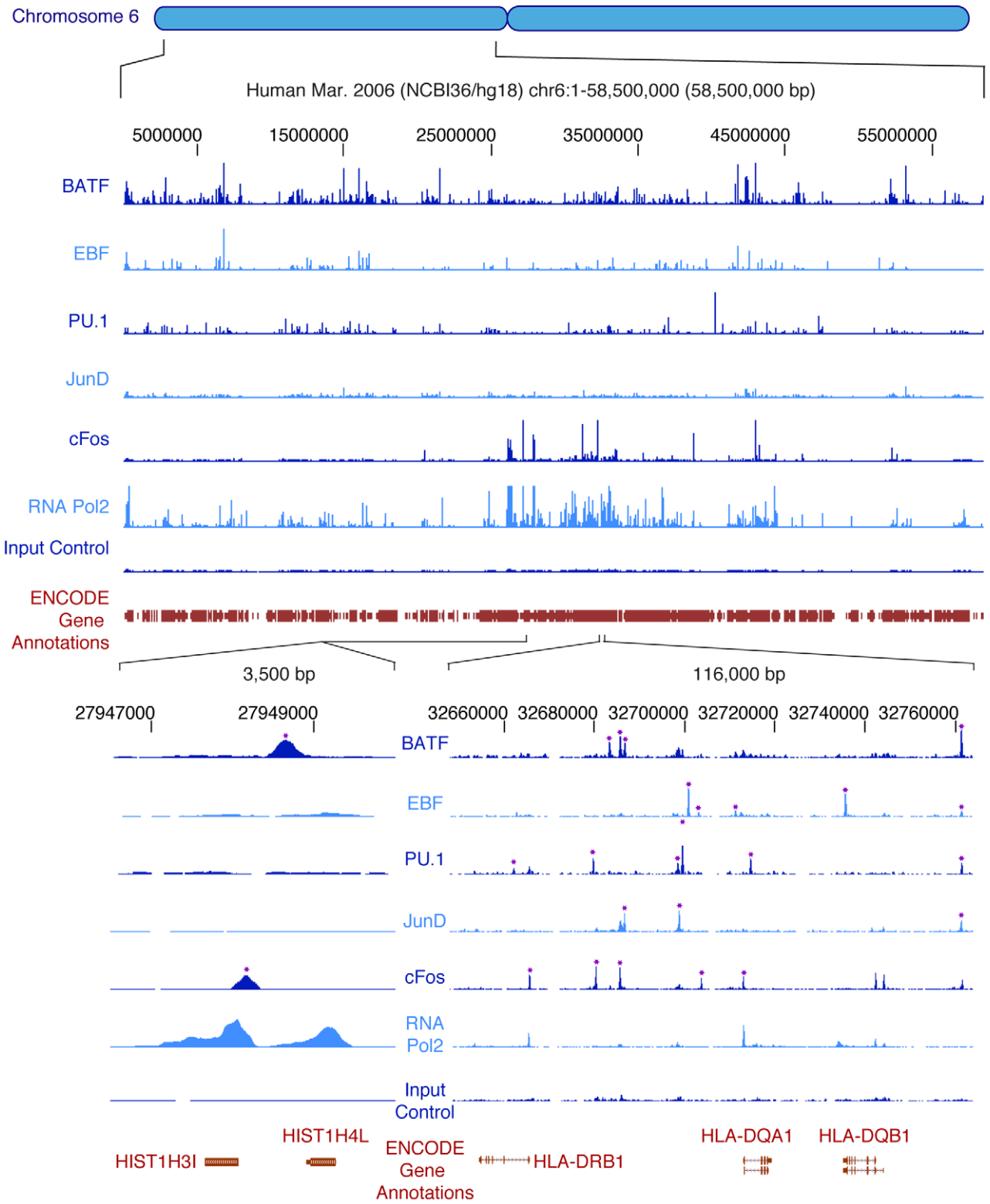

Figure 5. Occupancy of transcription factors and RNA polymerase 2 on human chromosome 6p as determined by ChIP-seq. The upper portion shows the ChIP-seq signal of five sequence-specific transcription factors and RNA Pol2 throughout the 58.5 Mb of the short arm of human chromosome 6 of the human lymphoblastoid cell line GM12878. Input control signal is shown below the RNA Pol2 data. At this level of resolution, the sites of strongest signal appear as vertical spikes in blue next to the name of each experiment (''BATF,'' ''EBF,'' etc.). More detail can be seen in the bottom right portion, where a 116 kb segment of the HLA region is expanded; here, individual sites of occupancy can be seen mapping to specific regions of the three HLA genes shown at the bottom, with asterisks indicating binding sites called by peak calling software. Finally, the lower left region shows a 3,500 bp region around two tandem histone genes, with RNA Pol2 occupancy at both promoters and two of the five transcription factors, BATF and cFos, occupying sites nearby. Selected annotations from the ENCODE Gene Annotations are shown in each case.
doi:10.1371/journal.pbio.1001046.g005

platform [79], and the results are reported as genotypes and as intensity value ratios for each allele. The genotype and sequence data from GM12878 generated by the 1,000 Genomes Project are being integrated with sequence data from ENCODE chromatin, transcription, TF occupancy, DNA methylation, and other assays to facilitate recognition of functional allelic variation, a significant contributor to phenotypic variability in gene expression [80,81]. The data also permit determination of the sequence copy number gains and losses found in every human genome [82–84], which are particularly prevalent in cell lines of malignant origin.

Long-range Chromatin interactions. Because *cis*-regulatory elements such as enhancers can control genes from distances of tens to hundreds of kb through looping interactions [85], a major challenge presented by ENCODE data is to connect distal regulatory elements with their cognate promoter(s). To map this connectivity, the Consortium is applying the 5C method [86], an enhanced version of Chromosome Conformation Capture (3C) [87], to selected cell lines. 5C has been applied comprehensively to the ENCODE pilot regions as well as to map the interactions between distal DNaseI hypersensitive sites and transcriptional start sites across chromosome 21 and selected domains throughout the genome. Special interfaces have been developed to visualize these 3-dimensional genomic data and are publicly available at http://my5C.umassmed.edu [88].

Protein:RNA interactions. RNA-binding proteins play a major role in regulating gene expression through control of mRNA translation, stability, and/or localization. Occupancy of RNA-binding proteins (RBPs) on RNA can be determined by using immunoprecipitation-based approaches (RIP-chip and RIP-seq) [89–92] analogous to those used for measuring TF occupancy. To generate maps of RBP:RNA associations and binding sites, a combination of RIP-chip and RIP-seq are being used. These approaches are currently targeting 4–6 RBPs in five human cell types (K562, GM12878, H1 ES, HeLa, and HepG2). RBP associations with non-coding RNA and with mRNA are also being explored.

Identification of functional elements with integrative analysis and fine-scale assays of biochemical elements. ChIP-seq of TFs and chromatin modifications may identify genomic regions bound by transcription factors in living cells but do not reveal which segments bound by a given TF are functionally important for transcription. By applying integrative approaches that incorporate histone modifications typical of enhancers (e.g., histone H3, Lysine 4 monomethylation), promoters (e.g., histone H3, Lysine 4 trimethylation), and silencers (e.g., Histone H3, Lysine 27, and Lysine 9 trimethylation), ENCODE is categorizing putative functional elements and testing a subset for activities in the context of transient transfection/reporter gene assays [93–97]. To further pinpoint the biological activities associated with specific regions of TF binding and chromatin modification within promoters, hundreds of TF binding sites have been mutagenized, and the mutant promoters are being assayed for effects on reporter gene transcription by transient transfection assays. This approach is enabling identification of specific TF binding sites that lead to activation and others associated with transcriptional repression.

Proteomics. To assess putative protein products generated from novel RNA transcripts and isoforms, proteins are sequenced and quantified by mass spectrometry and mapped back to their encoding transcripts [33,34,98]. ENCODE has recently begun to study proteins from distinct subcellular compartments of K562 and GM12878 with this complementary approach.

Evolutionary conservation. Evolutionary conservation is an important indicator of biological function. ENCODE is approaching evolutionary analysis from two directions. Functional properties are being assigned to conserved sequence elements identified through multi-species alignments, and conversely, the evolutionary histories of biochemically defined elements are being deduced. Multiple alignments of the genomes of 33 mammalian species have been constructed by using the Enredo, Pecan, Ortheus approach (EPO) [99,100], and complementary multiple alignments are available through the UCSC browser (UCSC Lastz/ChainNet/Multiz). These alignments enable measurement of evolutionary constraint at single-nucleotide resolution using GERP [101], SCONE [102], PhyloP [103], and other algorithms. In addition, conservation of DNA secondary structure based on hydroxyl radical cleavage patterns is being analyzed with the Chai algorithm [7].

Data Production Standards and Assessment of Data Quality

With the aim of ensuring quality and consistency, ENCODE has defined standards for collecting and processing each data type. These standards encompass all major experimental components, including cell growth conditions, antibody characterization, requirements for controls and biological replicates, and assessment of reproducibility. Standard formats for data submission are used that capture all relevant data parameters and experimental conditions, and these are available at the public ENCODE portal (http://genome.ucsc.edu/ENCODE/dataStandards.html). All ENCODE data are reviewed by a dedicated quality assurance team at the Data Coordination Center before release to the public. Experiments are considered to be *verified* when two highly concordant biological replicates have been obtained with the same experimental technique. In addition, a key quality goal of ENCODE is to provide *validation* at multiple levels, which can be further buttressed by cross-correlation between disparate data types. For example, we routinely perform parallel analysis of the same biological samples with alternate detection technologies (for example, ChIP-seq versus ChIP-chip or ChIP-qPCR). We have also compared our genome-wide results to "gold-standard" data from individual locus studies, such as DNase-seq versus independently performed conventional (Southern-based) DNaseI hypersensitivity studies. Cross-correlation of independent but related ENCODE data types with one another, such as DNaseI hypersensitivity, FAIRE, transcription factor occupancy, and histone modification patterns, can provide added confidence in the identification of specific DNA elements. Similarly, cross-correlation between long RNA-seq, CAGE, and TAF1 ChIP-seq data can strengthen confidence in a candidate location for transcription initiation. Finally, ENCODE is performing pilot tests for the biological activity of DNA elements to the predictive potential of various ENCODE biochemical signatures for certain biological functions. Examples include transfection assays in cultured human cells and injection assays in fish embryos to test for enhancer, silencer, or insulator activities in DNA elements identified by binding of specific groups of TFs or the presence of DNaseI hypersensitive sites or certain chromatin marks. Ultimately, defining the full biological role of a DNA element in its native chromosomal location and organismic context is the greatest challenge. ENCODE is beginning to approach this by integrating its data with results from other studies of in situ knockouts and/or knockdowns, or the identification of specific naturally occurring single base mutations and small deletions associated with changes in gene expression. However, we expect that deep insights into the function of most elements will ultimately come from the community of biologists who will build on ENCODE data or use them to complement their own experiments.

Current Scope and Completeness of ENCODE Data

A catalog of ENCODE datasets is available at http://encodeproject.org. These data provide evidence that ~1 Gigabase (Gb; 32%) of the human genome sequence is represented in steady-state, predominantly processed RNA populations. We have also delineated more than 2 million potential regulatory DNA regions through chromatin and TF mapping studies.

The assessment of the completeness of detection of any given element is challenging. To analyze the detection of transcripts in a single experiment, we have sequenced to substantial depth and used a sampling approach to estimate the number of reads needed to approach complete sampling of the RNA population (Figure 6A) [104]. For example, analyzing RNA transcripts with about 80 million mapped reads yields robust quantification of more than 80% of the lowest abundance class of genes (2–19 reads per kilobase per million mapped tags, RPKM) [24]. Measuring RNAs across multiple cell types, we find that, after the analysis of seven cell lines, 68% of the GENCODE transcripts can be detected with RPKM >1.

In the case of regulatory DNA, we have analyzed the detection of regulatory DNA by using three approaches: 1) the saturation of occupancy site discovery for a single transcription factor within a single cell type as a function of sequencing read depth, 2) the incremental discovery of DNaseI hypersensitive sites or the occupancy sites for a single TF across multiple cell types, and 3) the incremental rate of collective TF occupancy site discovery for all TFs across multiple cell types.

For detecting TF binding sites by ChIP-seq, we have found that the number of significant binding sites increases as a function of sequencing depth and that this number varies widely by transcription factor. For example, as shown in Figure 6B, 90% of detectable sites for the transcription factor GABP can be identified by using the MACS peak calling program at a depth of 24 million reads, whereas only 55% of detectable RNA Pol2 sites are identified at this depth when an antibody that recognizes both initiating and elongating forms of the enzyme is used. Even at 50 million reads, the number of sites is not saturated for RNA Pol2 with this antibody. It is important to note that determinations of saturation may vary with the use of different antibodies and laboratory protocols. For instance, a different RNA Pol2 antibody that recognizes unphosphorylated, non-elongating RNA Pol2 bound only at promoters requires fewer reads to reach saturation [105]. For practical purposes, ENCODE currently uses a minimum sequencing depth of 20 M uniquely mapped reads for sequence-specific transcription factors. For data generated prior to June 1, 2010, this figure was 12 M.

To assess the incremental discovery of regulatory DNA across different cell types, it was necessary to account for the non-uniform correlation between cell lines and assays (see Figure 6C legend for details). We therefore examined all possible orderings of either cell types or assays and calculated the distribution of elements discovered as the number of cell types or assays increases, presented as saturation distribution plots (Figure 6C and 6D, respectively). For DNase hypersensitive sites, we observe a steady increase in the mean number of sites discovered as additional cell types are tested up to and including the 62 different cell types examined to date, indicating that new elements continue to be identified at a relatively high rate as additional cell types are sampled (Figure 6C). Analysis of CTCF sites across 28 cell types using this approach shows similar behavior. Analysis of binding sites for 42 TFs in the cell line with most data (K562) also shows that saturation of the binding sites for these factors has not yet been achieved. These results indicate that additional cell lines need to be analyzed for DNaseI and many transcription factors, and that many more transcription factors need to be analyzed within single cell types to capture all the regulatory information for a given factor across the genome. The implications of these trends for defining the extent of regulatory DNA within the human genome sequence is as yet unclear.

III. Accessing ENCODE Data

ENCODE Data Release and Use Policy

The ENCODE Data Release and Use Policy is described at http://www.encodeproject.org/ENCODE/terms.html. Briefly, ENCODE data are released for viewing in a publicly accessible browser (initially at http://genome-preview.ucsc.edu/ENCODE and, after additional quality checks, at http://encodeproject.org). The data are available for download and pre-publication analysis of any kind, as soon as they are verified (i.e., shown to be reproducible). However, consistent with the principles stated in the Toronto Genomic Data Use Agreement [106], the ENCODE Consortium data producers request that they have the first publication on genome-wide analyses of ENCODE data, within a 9-month timeline from its submission. The timeline for each dataset is clearly displayed in the information section for each dataset. This parallels policies of other large consortia, such as the HapMap Project (http://www.hapmap.org), that attempt to balance the goal of rapid data release with the ability of data producers to publish initial analyses of their work. Once a producer has published a dataset during this 9-month period, anyone may publish freely on the data. The embargo applies only to global analysis, and the ENCODE Consortium expects and encourages immediate use and publication of information at one or a few loci, without any consultation or permission. For such uses, identifying ENCODE as the source of the data by citing this article is requested.

Public Repositories of ENCODE Data

After curation and review at the Data Coordination Center, all processed ENCODE data are publicly released to the UCSC Genome Browser database (http://genome.ucsc.edu). Accessioning of ENCODE data at the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/info/ENCODE.html) is underway. Primary DNA sequence reads are stored at UCSC and the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?) and will also be retrievable via GEO. Primary data derived from DNA microarrays (for example, for gene expression) are deposited directly to GEO. The processed data are also formatted for viewing in the UCSC browser. Metadata, including information on antibodies, cell culture conditions, and other experimental parameters, are deposited into the UCSC database, as are results of validation experiments. Easy retrieval of ENCODE data to a user's desktop is facilitated by the UCSC Table Browser tool (http://genome.ucsc.edu/cgi-bin/hgTables?org=human), which does not require programming skills. Computationally sophisticated users may gain direct access to data through application programming interfaces (APIs) at both the UCSC browser and NCBI and by downloading files from http://genome.ucsc.edu/ENCODE/downloads.html.

An overview of ENCODE data types and the location of the data repository for each type is presented in Table 4.

IV. Working with ENCODE Data

Using ENCODE Data in the UCSC Browser

Many users will want to view and interpret the ENCODE data for particular genes of interest. At the online ENCODE portal

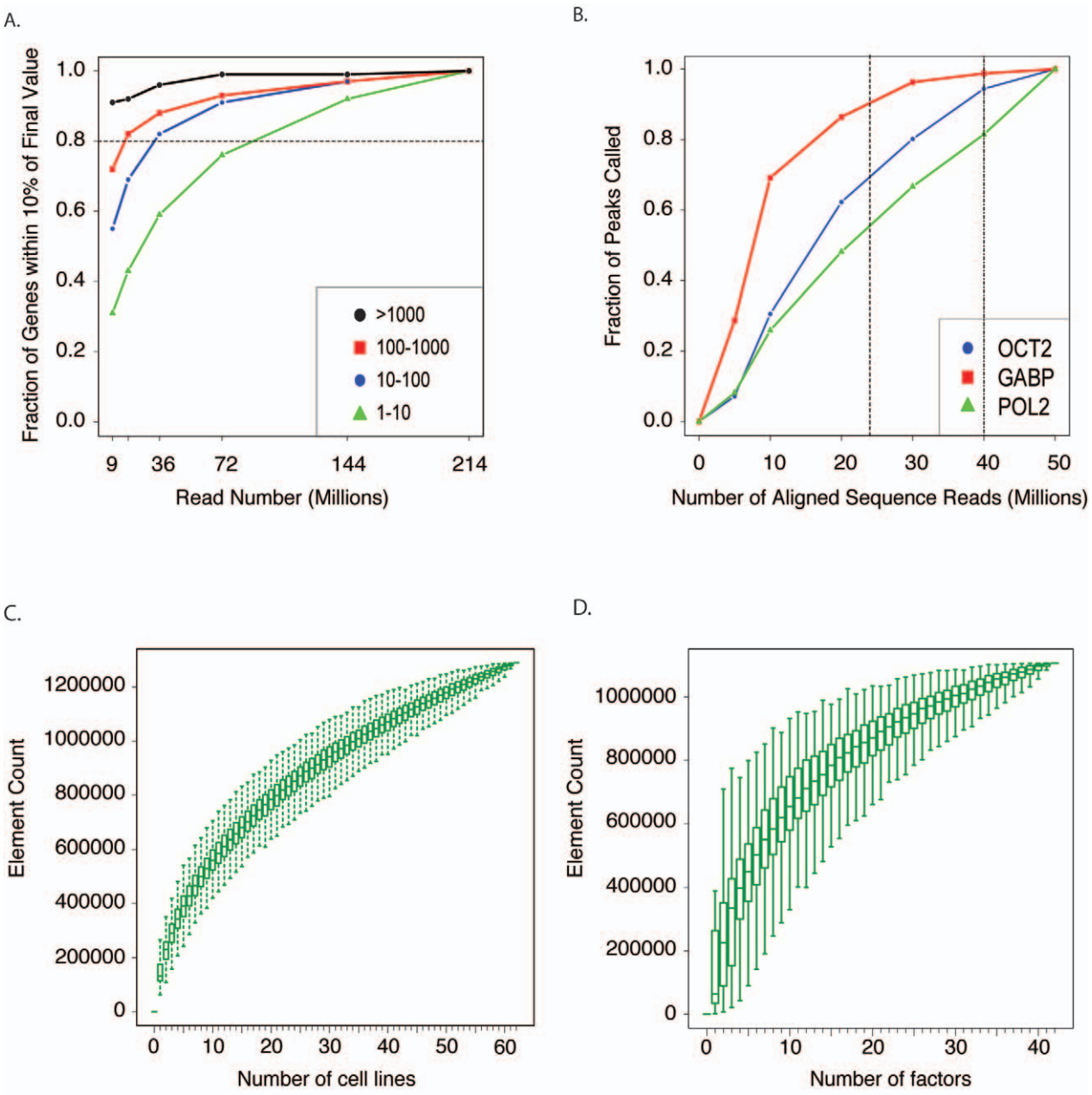

Figure 6. Incremental discovery of transcribed elements and regulatory DNA. (A) Robustness of gene expression quantification relative to sequencing depth. PolyA-selected RNA from H1 human embryonic stem cells was sequenced to 214 million mapped reads. The number of reads (indicated on the *x*-axis) was sampled from the total, and gene expression (in FPKM) was calculated and compared to the gene expression values resulting from all the reads (final values). Gene expression levels were split into four abundance classes and the fraction of genes in each class with RPKM values within 10% of the final values was calculated. At ~80 million mapped reads, more than 80% of the low abundance class of genes is robustly quantified according to this measure (horizontal dotted line). Abundances for the classes in RPKM are given in the inset box. (B) Effect of number of reads on fractions of peaks called in ChIP-seq. ChIP-seq experiments for three sequence-specific transcription factors were sequenced to a depth of 50 million aligned reads. To evaluate the effect of read depth on the number of binding sites identified, peaks were called with the MACS algorithm at various read depths, and the fraction of the total number of peaks that were identified at each read depth are shown. For sequence-specific transcription factors that have strong signal with ChIP-seq, such as GABP, approximately 24 million reads (dashed vertical line) are sufficient to capture 90% of the binding sites. However, for more general sequence-specific factors (e.g., OCT2), additional sequencing continues to yield additional binding site information. RNA Pol2, which interacts with DNA broadly across genes, maintains a nearly linear gain in binding information through 50 million aligned reads. (C) Saturation analysis of ENCODE DNaseI hypersensitivity data with increasing numbers of cell lines. The plot shows the extent of saturation of DNaseI hypersensitivity sites (DHSs) discovered as increasing numbers of cell lines are studied. The plot is generated from the ENCODE DNaseI elements defined at the end of January 2010 (from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC) as follows. We first define a set of DHSs from the overlap of all DHS data across all cell lines. Where overlapping elements are identified in two or more cell lines, these are determined to represent the same element and fused up to a maximum size of 5 kb. Elements above this limit are split and counted as distinct. We then calculate the subset of these elements represented by each single cell line experiment. The distribution of element counts for each single cell line is plotted as a box plot with the median at position 1 on the *x*-axis. We next calculate the element contributions of all possible pairs of cell line experiments and plot this distribution at position 2. We continue to do this for all incremental steps up to and including all cell lines (which is

by definition only a single data point). (D) Saturation of TF ChIP-seq elements in K562 cells. This plot illustrates the saturation of elements identified by TF ChIP-seq as additional factors are analyzed within the same cell line. The plot is generated by the equivalent approach as described in (C), except the data are now the set of all elements defined by ChIP-seq analysis of K562 cells with 42 different transcription factors. The data were from the January 2010 data freeze from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC. For consistency, the peak calls from all ChIP-seq data were generated by a uniform processing pipeline with the Peakseq peak caller and IDR replicate reconciliation.
doi:10.1371/journal.pbio.1001046.g006

(http://encodeproject.org), users should follow a "Genome Browser" link to visualize the data in the context of other genome annotations. Currently, it is useful for users to examine both the hg18 and the hg19 genome browsers. The hg18 has the ENCODE Integrated Regulation Track on by default, which shows a huge amount of data in a small amount of space. The hg19 browser has newer datasets, and more ENCODE data than are available on hg18. Work is in progress to remap the older hg18 datasets to hg19 and generate integrated ENCODE tracks. On either browser, additional ENCODE tracks are marked by a double helix logo in the browser track groups for genes, transcripts, and regulatory features. Users can turn tracks on or off to develop the views most useful to them (Figure 7). To aid users in navigating the rich variety of data tracks, the ENCODE portal also provides a detailed online tutorial that covers data display, data download, and analysis functions available through the browser. Examples applying ENCODE data at individual loci to specific biological or medical issues are a good starting point for exploration and use of the data. Thus, we also provide a collection of examples at the "session gallery" at the ENCODE portal. Users are encouraged to submit additional examples; we anticipate that this community-based sharing of insights will accelerate the use and impact of the ENCODE data.

An Illustrative Example

Numerous genome-wide association studies (GWAS) that link human genome sequence variants with the risk of disease or with common quantitative phenotypes have now become available. However, in most cases, the molecular consequences of disease- or trait-associated variants for human physiology are not understood [107]. In more than 400 studies compiled in the GWAS catalog [108], only a small minority of the trait/disease-associated SNPs (TASs) occur in protein-coding regions; the large majority (89%) are in noncoding regions. We therefore expect that the accumulating functional annotation of the genome by ENCODE will contribute substantially to functional interpretation of these TASs.

For example, common variants within a ~1 Mb region upstream of the c-Myc proto-oncogene at 8q24 have been associated with cancers of the colon, prostate, and breast (Figure 8A) [109–111]. ENCODE data on transcripts, histone modifications, DNase hypersensitive sites, and TF occupancy show strong, localized signals in the vicinity of major cancer-associated SNPs. One variant (*rs698327*) lies within a DNase hypersensitive site that is bound by several TFs and the enhancer-associated protein p300 and contains histone modification patterns typical of enhancers (high H3K4me1, low H3K4me3; Figure 8B). Recent studies have shown enhancer activity and allele-specific binding of TCF7L2 at this site [112], with the risk allele showing greater binding and activity [113,114]. Moreover, this element appears to contact the downstream c-Myc gene in vivo, compatible with enhancer function [114,115]. Similarly, several regions predicted via ENCODE data to be involved in gene regulation are close to SNPs in the *BCL11A* gene associated with persistent expression of fetal hemoglobin (Figure S2). These examples show that the simple overlay of ENCODE data with candidate non-coding risk-associated variants may readily identify specific genomic elements as leading candidates for investigation as probable effectors of phenotypic effects via alterations in gene expression or other genomic regulatory processes. Importantly, even data from cell types not directly associated with the phenotype of interest may be of considerable value for hypothesis generation. It is reasonable to expect that application of current and future ENCODE data will provide useful information concerning the mechanism(s) whereby genomic variation influences susceptibility to disease, which then can then be tested experimentally.

Limitations of ENCODE Annotations

All ENCODE datasets to date are from populations of cells. Therefore, the resulting data integrate over the entire cell population, which may be physiologically and genetically inhomogeneous. Thus, the source cell cultures in the ENCODE experiments are not typically synchronized with respect to the cell cycle and, as with all such samples, local micro-environments in culture may also vary, leading to physiological differences in cell state within each culture. In addition, one Tier 1 cell line (K562) and two Tier 2 cell lines (HepG2 and HeLa) are known to have abnormal genomes and karyotypes, with genome instability. Finally, some future Tier 3 tissue samples or primary cultures may be inherently heterogeneous in cell type composition. Averaging over heterogeneity in physiology and/or genotype produces an amalgamation of the contributing patterns of gene

Table 4. Overview of ENCODE data types.

Data	Description	Location
Metadata	Experimental parameters (e.g., growth conditions, antibody characterization)	UCSC, GEO
Primary data images	CCD camera images from sequencers or microarrays	Not archived
Sequence reads/microarray signal	Minimally processed experimental data; reads and quality information; probe locations and intensities	UCSC, GEO, SRA
Aligned sequence reads	Sequence reads and genomic positions	UCSC, GEO
Genomic signal	Sequence tag density (sliding window); cumulative base coverage or density by sequencing or read pseudo-extension; microarray probe intensity	UCSC, GEO
Enriched region calls/scores/*p* or *q* values	Putative binding or transcribed regions	UCSC, GEO

doi:10.1371/journal.pbio.1001046.t004

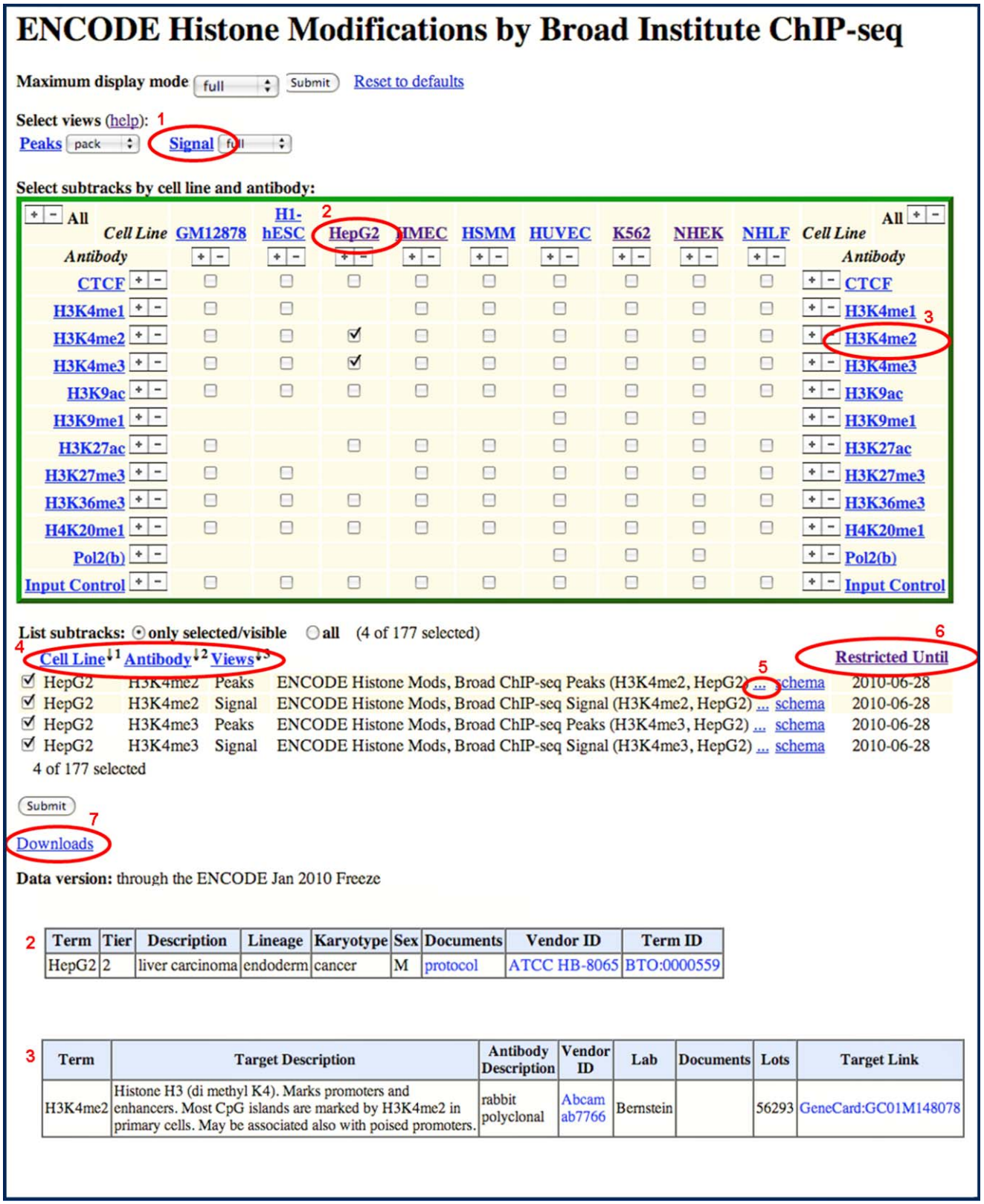

Figure 7. Accessing ENCODE data at the UCSC Portal. Data and results for the ENCODE Project are accessible at the UCSC portal (http://genome.ucsc.edu/ENCODE). "Signal tracks" for the different datasets are selected and displayed in the genome browser to generate images such as those shown in Figures 3–4. The datasets are available from the Track Settings page; an example is shown that illustrates some of the key controls. A dataset is selected and the Signal display plots the values of an assay for a given feature more or less continuously along a chromosome. The height, range for the *y*-axis, windowing function, and many other aspects of the graph are controlled in the Signal Configuration window, accessed by clicking on "Signal" (red oval #1). ENCODE data are commonly generated on multiple cell lines; information about each can be accessed by clicking

on the name of the cell line or antibody (e.g., HepG2, red oval #2). Many ENCODE tracks are actually composites of multiple subtracks; these can be turned on and off by using the boxes in the central matrix or in the subtrack list below. Subtracks can be reordered individually by using drag and drop in the browser image or the Track Settings page, or in logical groups by using the "Cell/Antibody/Views" (red oval #4) ordering controls. Additional information about the feature and the assay, such as the antibody used, can be obtained by clicking on the name of the feature. Some restrictions to the use of ENCODE data apply for a 9-month period after deposit of the data; the end of that 9-month period is given by the "Restricted Until" date. Full data can be downloaded by clicking on the "Downloads" link (red oval #7).
doi:10.1371/journal.pbio.1001046.g007

expression, factor occupancy, and chromatin status that must be considered when using the data. Future improvements in genome-wide methodology that allow the use of much smaller amounts of primary samples, or follow-up experiments in single cells when possible, may allow us to overcome many of these caveats.

The use of DNA sequencing to annotate functional genomic features is constrained by the ability to place short sequence reads accurately within the human genome sequence. Most ENCODE data types currently represented in the UCSC browser use only those sequence reads that map uniquely to the genome. Thus, centromeric and telomeric segments (collectively ~15% of the genome and enriched in recent transposon insertions and segmental duplications) as well as sequences not present in the current genome sequence build [116] are not subject to reliable annotation by our current techniques. However, such information can be gleaned through mining of the publicly available raw sequence read datasets generated by ENCODE.

It is useful to recognize that the confidence with which different classes of ENCODE elements can be related to a candidate function varies. For example, ENCODE can identify with high confidence new internal exons of protein-coding genes, based on RNA-seq data for long polyA+ RNA. Other features, such as candidate promoters, can be identified with less, yet still good, confidence by combining data from RNA-seq, CAGE-tags, and RNA polymerase 2 (RNA Pol2) and TAF1 occupancy. Still other ENCODE biochemical signatures come with much lower confidence about function, such as a candidate transcriptional enhancer supported by ChIP-seq evidence for binding of a single transcription factor.

Identification of genomic regions enriched by ENCODE biochemical assays relies on the application of statistical analyses and the selection of threshold significance levels, which may vary between the algorithms used for particular data types. Accordingly, discrete annotations, such as TF occupancy or DNaseI hypersensitive sites, should be considered in the context of reported p values, q values, or false discovery rates, which are conservative in many cases. For data types that lack focal enrichment, such as certain histone modifications and many RNA Pol2-bound regions, broad segments of significant enrichment have been delineated that encompass considerable quantitative variation in the signal strength along the genome.

V. ENCODE Data Analysis

Development and implementation of algorithms and pipelines for processing and analyzing data has been a major activity of the ENCODE Project. Because massively parallel DNA sequencing has been the main type of data generated by the Consortium, much of the algorithmic development and data analysis to date has been concerned with issues related to producing and interpreting such data. Software packages and algorithms commonly used in the ENCODE Consortium are summarized in Tables 3 and S1.

In general, the analysis of sequencing-based measurements of functional or biochemical genomic parameters proceeds through three major phases. In the first phase, the short sequences that are the output of the experimental method are aligned to the reference genome. Algorithm development for efficient and accurate alignment of short read sequences to the human genome is a rapidly developing field, and ENCODE groups employ a variety of the state-of-the-art software (see Tables 3 and S1). In the second phase, the initial sequence mapping is processed to identify significantly enriched regions from the read density. For ChIP-seq (TFs and histone modification), DNase-seq or FAIRE-seq, both highly localized peaks or broader enriched regions may be identified. Within the ENCODE Consortium, each data production group provides lists of enriched regions or elements within their own data, which are available through the ENCODE portal. It should be noted that, for most data types, the majority of enriched regions show relatively weak absolute signal, necessitating the application of conservative statistical thresholds. For some data, such as those derived from sampling RNA species (e.g., RNA-seq), additional algorithms and processing are used to handle transcript structures and the recognition of splicing events.

The final stage of analysis involves integrating the identified regions of enriched signal with each other and with other data types. An important prerequisite to data integration is the availability of uniformly processed datasets. Therefore, in addition to the processing pipelines developed by individual production groups, ENCODE has devoted considerable effort toward establishing robust uniform processing for phases 1 and 2 to enable integration. For signal comparison, specific consideration has been given to deriving a normalized view of the sequence read density of each experiment. In the case of ChIP-seq for TFs, this process includes in silico extension of the sequence alignment to reflect the experimentally determined average lengths of the input DNA molecules that are sampled by the short sequence tag, compensation for repetitive sequences that may lead to alignment with multiple genomic locations, and consideration of the read density of the relevant control or input chromatin experiment. ENCODE has adopted a uniform standardized peak-calling approach for transcription factor ChIP-seq, including a robust and conservative replicate reconciliation statistic (Irreproducible Discovery Rate, IDR [117], to yield comparable consensus peak calls. As the project continues, we expect further standardizations to be developed.

There are many different ways to analyze and integrate large, diverse datasets. Some of the basic approaches include assigning features to existing annotations (e.g., assigning transcribed regions to annotated genes or Pol2-binding peaks to likely genes), discovery of correlations among features, and identification of particular gene classes (e.g., Gene Ontology categories) preferentially highlighted by a given annotation. Many software tools exist in the community for these purposes, including some developed within the ENCODE Project, such as the Genome Structure Correction statistic for assessing overlap significance [3]. Software tools used for integration by ENCODE are summarized in Tables 3 and S1.

VI. Future Plans and Challenges

Data Production Plans

The challenge of achieving complete coverage of all functional elements in the human genome is substantial. The adult human body contains several hundred distinct cell types, each of which

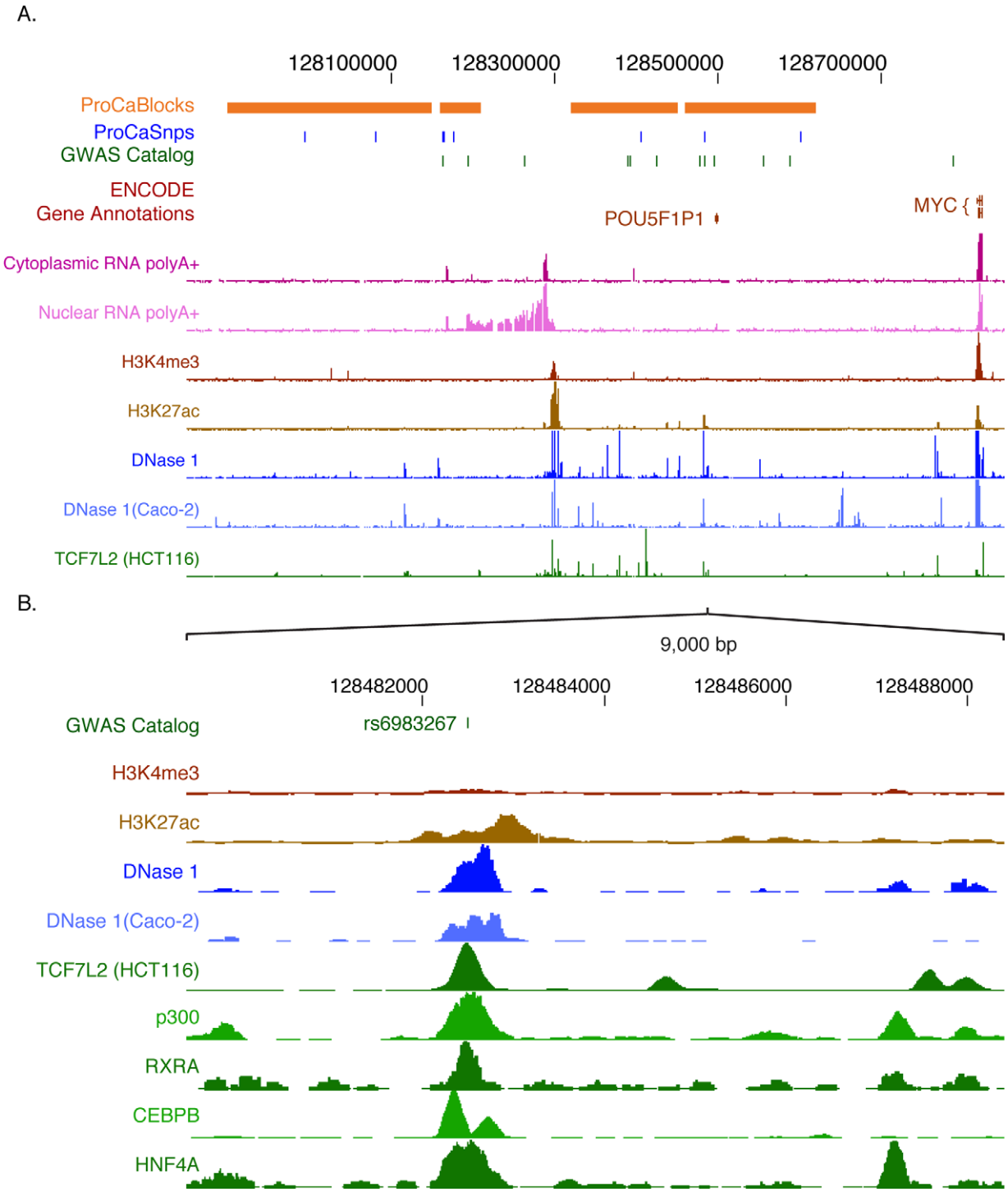

Figure 8. ENCODE data indicate non-coding regions in the human chromosome 8q24 loci associated with cancer. (A) A 1 Mb region including *MYC* and a gene desert upstream shows the linkage disequilibrium blocks and positions of SNPs associated with breast and prostate cancer, with both a custom track based on [121] and the resident track from the GWAS catalog. ENCODE tracks include GENCODE gene annotations, results of mapping RNAs to high-density Affymetrix tiling arrays (cytoplasmic and nuclear polyA+ RNA), mapping of histone modifications (H3K4me3 and H3K27Ac), DNaseI hypersensitive sites in liver and colon carcinoma cell lines (HepG2 and Caco-2), and occupancy by the transcription factor TCF7L2 in HCT116 cells. (B) Expanded view of a 9 kb region containing the cancer-associated SNP *rs6983267* (shown on the top line). In addition to the histone modifications, DNaseI hypersensitive sites and factor occupancy described in (A), the ENCODE tracks also show occupancy by the coactivator p300 and the transcription factors RXRA, CEBPB, and HNF4A. Except as otherwise noted in brackets, the ENCODE data shown here are from the liver carcinoma cell line HepG2.
doi:10.1371/journal.pbio.1001046.g008

expresses a unique subset of the ~1,500 TFs encoded in the human genome [118]. Furthermore, the brain alone contains thousands of types of neurons that are likely to express not only different sets of TFs but also a larger variety of non-coding RNAs [119]. In addition, each cell type may exhibit a diverse array of responses to exogenous stimuli such as environmental conditions or chemical agents. Broad areas of fundamental chromosome function, such as meiosis and recombination, remain unexplored. Furthermore, ENCODE has focused chiefly on definitive cells and cell lines, bypassing the substantial complexity of development and differentiation. A truly comprehensive atlas of human functional elements is not practical with current technologies, motivating our focus on performing the available assays in a range of cell types that will provide substantial near-term utility. ENCODE is currently developing a strategy for addressing this cellular space in a timely manner that maximizes the value to the scientific community. Feedback from the user community will be a critical component of this process.

Integrating ENCODE with Other Projects and the Scientific Community

To understand better and functionally annotate the human genome, ENCODE is making efforts to analyze and integrate data within the project and with other large-scale projects. These efforts include 1) defining promoter and enhancer regions by combining transcript mapping and biochemical marks, 2) delineating distinct classes of regions within the genomic landscape by their specific combinations of biochemical and functional characteristics, and 3) defining transcription factor co-associations and regulatory networks. These efforts aim to extend our understanding of the functions of the different biochemical elements in gene regulation and gene expression.

One of the major motivations for the ENCODE Project has been to aid in the interpretation of human genome variation that is associated with disease or quantitative phenotypes. The Consortium is therefore working to combine ENCODE data with those from other large-scale studies, including the 1,000 Genomes Project, to study, for example, how SNPs and structural variation may affect transcript, regulatory, and DNA methylation data. We foresee a time in the near future when the biochemical features defined by ENCODE are routinely combined with GWAS and other sequence variation–driven studies of human phenotypes. Analogously, the systematic profiling of epigenomic features across ex vivo tissues and stem cells currently being undertaken by the NIH Roadmap Epigenomics program will provide synergistic data and the opportunity to observe the state and behavior of ENCODE-identified elements in human tissues representing healthy and disease states.

These are but a few of many applications of the ENCODE data. Investigators focused on one or a few genes should find many new insights within the ENCODE data. Indeed, these investigators are in the best position to infer potential functions and mechanisms from the ENCODE data—ones that will also lead to testable hypotheses. Thus, we expect that the work of many investigators will be enhanced by these data and that their results will in turn inform the development of the project going forward.

Finally, we also expect that comprehensive paradigms for gene regulation will begin to emerge from our work and similar work from many laboratories. Deciphering the "regulatory code" within the genome and its associated epigenetic signals is a grand and complex challenge. The data contributed by ENCODE in conjunction with complementary efforts will be foundational to this effort, but equally important will be novel methods for genome-wide analysis, model building, and hypothesis testing. We therefore expect the ENCODE Project to be a major contributor not only of data but also novel technologies for deciphering the human genome and those of other organisms.

Supporting Information

Figure S1 The Organization of the ENCODE Consortium. The geographical distribution of the members of the ENCODE Consortium, with pin colors indicating the group roles as detailed in the text below.
(TIF)

Figure S2 Quantitative trait example (BCL11A). Candidates for gene regulatory features in the vicinity of SNPs at the *BCL11A* locus associated with fetal hemoglobin levels. SNPs associated with fetal hemoglobin levels are marked in red on the top line; those not associated are marked in blue. The phenotype-associated SNPs are close to an antisense transcript (AC009970.1, light orange), shown in the ENCODE gene annotations. This antisense transcript is within a region (boxed in red) with elevated levels of H3K 4me1 and DNase hypersensitive sites. The phenotype-associated region is flanked by two regions (boxed in blue) with multiple strong biochemical signals associated with transcriptional regulation, including transcription factor occupancy. The data are from the lymphoblastoid cell line GM12878, as *BCL11A* is expressed in this cell line (RNA-seq track) but not in K562 (unpublished data).
(TIF)

Table S1 This supplemental table contains additional details of the computational analysis tools used by the ENCODE Consortium that are listed in Table 3. The name of each software tool appears in the first column, and subsequent columns contain the tasks for which the tool is used, the PMID reference number when available, and a web address where the tool can be accessed.
(DOC)

Acknowledgments

We thank Judy R. Wexler and Julia Zhang at the National Human Genome Research Institute for their support in administering the ENCODE Consortium, additional members of our laboratories and institutions who have contributed to the experimental and analytical components of this project, and J. D. Frey for assistance in preparing the figures.

The ENCODE Consortium Authors

Writing Group. Richard M. Myers[1], John Stamatoyannopoulos[2], Michael Snyder[3], Ian Dunham[4], Ross C. Hardison[5], Bradley E. Bernstein[6,7], Thomas R. Gingeras[8], W. James Kent[9], Ewan Birney[4], Barbara Wold[10,11], Gregory E. Crawford[12,13].

Broad Institute Group. Bradley E. Bernstein[6,7], Charles B. Epstein[6], Noam Shoresh[6], Jason Ernst[6,14], Tarjei S. Mikkelsen[6], Pouya Kheradpour[6,14], Xiaolan Zhang[6], Li Wang[6], Robbyn Issner[6], Michael J. Coyne[6], Timothy Durham[6], Manching Ku[6], Thanh Truong[6], Lucas D. Ward[6,14], Robert C. Altshuler[14], Michael F. Lin[6,14], Manolis Kellis[6,14].

Cold Spring Harbor; University of Geneva; Center for Genomic Regulation, Barcelona; RIKEN; University of Lausanne; Genome Institute of Singapore Group. *Cold Spring Harbor I*: Thomas R. Gingeras[8], Carrie A. Davis[8], Philipp Kapranov[15], Alexander Dobin[8], Christopher Zaleski[8], Felix Schlesinger[8], Philippe Batut[8], Sudipto Chakrabortty[8], Sonali Jha[8], Wei Lin[8], Jorg Drenkow[8], Huaien Wang[8], Kim Bell[8], Hui Gao[16], Ian Bell[15], Erica Dumais[15], Jacqueline Dumais[15]. *University of Geneva*: Stylianos E. Antonarakis[17], Catherine Ucla[17], Christelle Borel[17]. *Center for Genomic Regulation, Barcelona*: Roderic Guigo[18], Sarah Djebali[18], Julien Lagarde[18], Colin Kingswood[18], Paolo Ribeca[18], Micha Sammeth[18], Tyler Alioto[18], Angelika Merkel[18], Hagen Tilgner[18]. *RIKEN*: Piero Carninci[19], Yoshihide Hayashizaki[19], Timo Lassmann[19], Hazuki Takahashi[19], Rehab F. Abdelhamid[19]. *Cold Spring Harbor II*: Gregory Hannon[20], Katalin Fejes-Toth[8], Jonathan Preall[8], Assaf Gordon[8], Vihra Sotirova[8]. *University of Lausanne*: Alexandre Reymond[21], Cedric Howald[21],

Emilie Aït Yahya Graison[21], Jacqueline Chrast[21]. *Genome Institute of Singapore*: Yijun Ruan[22], Xiaoan Ruan[22], Atif Shahab[22], Wan Ting Poh[22], Chia-Lin Wei[22].

Duke University, EBI, University of Texas, Austin, University of North Carolina–Chapel Hill Group. *Duke University*: Gregory E. Crawford[12,13], Terrence S. Furey[12], Alan P. Boyle[12], Nathan C. Sheffield[12], Lingyun Song[12], Yoichiro Shibata[12], Teresa Vales[12], Deborah Winter[12], Zhancheng Zhang[12], Darin London[12], Tianyuan Wang[12]. *EBI*: Ewan Birney[4], Damian Keefe[4]. *University of Texas, Austin*: Vishwanath R. Iyer[23], Bum-Kyu Lee[23], Ryan M. McDaniell[23], Zheng Liu[23], Anna Battenhouse[23], Akshay A. Bhinge[23]. *University of North Carolina–Chapel Hill*: Jason D. Lieb[24], Linda L. Grasfeder[24], Kimberly A. Showers[24], Paul G. Giresi[24], Seul K. C. Kim[24], Christopher Shestak[24].

HudsonAlpha Institute, Caltech, Stanford Group. *HudsonAlpha Institute*: Richard M. Myers[1], Florencia Pauli[1], Timothy E. Reddy[1], Jason Gertz[1], E. Christopher Partridge[1], Preti Jain[1], Rebekka O. Sprouse[1], Anita Bansal[1], Barbara Pusey[1], Michael A. Muratet[1], Katherine E. Varley[1], Kevin M. Bowling[1], Kimberly M. Newberry[1], Amy S. Nesmith[1], Jason A. Dilocker[1], Stephanie L. Parker[1], Lindsay L. Waite[1], Krista Thibeault[1], Kevin Roberts[1], Devin M. Absher[1]. *Caltech*: Barbara Wold[10,11], Ali Mortazavi[10,11], Brian Williams[10], Georgi Marinov[10], Diane Trout[10], Shirley Pepke[25], Brandon King[10], Kenneth McCue[10], Anthony Kirilusha[10], Gilberto DeSalvo[10], Katherine Fisher-Aylor[10], Henry Amrhein[10], Jost Vielmetter[11]. *Stanford*: Gavin Sherlock[3], Arend Sidow[3,26], Serafim Batzoglou[27], Rami Rauch[3], Anshul Kundaje[26,27], Max Libbrecht[27].

NHGRI Groups. *NHGRI, Genome Informatics Section*: Elliott H. Margulies[28], Stephen C. J. Parker[28]. *NHGRI, Genomic Functional Analysis Section*: Laura Elnitski[29]. *NHGRI, NIH Intramural Sequencing Center*: Eric D. Green[30].

Sanger Institute; Washington University; Yale University; Center for Genomic Regulation, Barcelona; UCSC; MIT; University of Lausanne; CNIO Group. *Sanger Institute*: Tim Hubbard[31], Jennifer Harrow[31], Stephen Searle[31], Felix Kokocinski[31], Browen Aken[31], Adam Frankish[31], Toby Hunt[31], Gloria Despacio-Reyes[31], Mike Kay[31], Gaurab Mukherjee[31], Alexandra Bignell[31], Gary Saunders[31], Veronika Boychenko[31]. *Washington University*: Michael Brent[32], M. J. Van Baren[32], Randall H. Brown[32]. *Yale University*: Mark Gerstein[33,34,35], Ekta Khurana[33,34], Suganthi Balasubramanian[33,34], Zhengdong Zhang[33,34], Hugo Lam[33,34], Philip Cayting[3,33,34], Rebecca Robilotto[33,34], Zhi Lu[33,34]. *Center for Genomic Regulation, Barcelona*: Roderic Guigo[18], Thomas Derrien[18], Andrea Tanzer[18], David G. Knowles[18], Marco Mariotti[18]. *UCSC*: W. James Kent[9], David Haussler[9,36], Rachel Harte[9], Mark Diekhans[9]. *MIT*: Manolis Kellis[6,14], Mike Lin[6,14], Pouya Kheradpour[6,14], Jason Ernst[6,14]. *University of Lausanne*: Alexandre Reymond[21], Cedric Howald[21], Emilie Aït Yahya Graison[21], Jacqueline Chrast[21]. *CNIO*: Alfonso Valencia[37], Michael Tress[37], Jose Manuel Rodriguez[37].

Stanford-Yale, Harvard, University of Massachusetts Medical School, University of Southern California/UCDavis Group. *Stanford-Yale*: Michael Snyder[3], Stephen G. Landt[3], Debasish Raha[38], Minyi Shi[3], Ghia Euskirchen[3], Fabian Grubert[3], Maya Kasowski[38], Jin Lian[39], Philip Cayting[3,33,34], Phil Lacroute[3], Youhan Xu[38], Hannah Monahan[38], Dorrelyn Patacsil[3], Teri Slifer[3], Xinqiong Yang[3], Alexandra Charos[38], Brian Reed[38], Linfeng Wu[3], Raymond K. Auerbach[33], Lukas Habegger[33], Manoj Hariharan[3], Joel Rozowsky[33,34], Alexej Abyzov[33,34], Sherman M. Weissman[39], Mark Gerstein[33,34,35]. *Harvard*: Kevin Struhl[40], Nathan Lamarre-Vincent[40], Marianne Lindahl-Allen[40], Benoit Miotto[40], Zarmik Moqtaderi[40], Joseph D. Fleming[40]. *University of Massachusetts Medical School*: Peter Newburger[41]. *University of Southern California/UCDavis*: Peggy J. Farnham[42,43], Seth Frietze[42,43], Henriette O'Geen[43], Xiaoqin Xu[43], Kim R. Blahnik[43], Alina R. Cao[43], Sushma Iyengar[43].

University of Washington, University of Massachusetts Medical School Group. *University of Washington*: John A. Stamatoyannopoulos[2], Rajinder Kaul[2], Robert E. Thurman[2], Hao Wang[2], Patrick A. Navas[2], Richard Sandstrom[2], Peter J. Sabo[2], Molly Weaver[2], Theresa Canfield[2], Kristen Lee[2], Shane Neph[2], Vaughan Roach[2], Alex Reynolds[2], Audra Johnson[2], Eric Rynes[2], Erika Giste[2], Shinny Vong[2], Jun Neri[2], Tristan Frum[2], Ericka M. Johnson[2], Eric D. Nguyen[2], Abigail K. Ebersol[2], Minerva E. Sanchez[2], Hadar H. Sheffer[2], Dimitra Lotakis[2], Eric Haugen[2], Richard Humbert[2], Tanya Kutyavin[2], Tony Shafer[2]. *University of Massachusetts Medical School*: Job Dekker[44], Bryan R. Lajoie[44], Amartya Sanyal[44].

Data Coordination Center. W. James Kent[9], Kate R. Rosenbloom[9], Timothy R. Dreszer[9], Brian J. Raney[9], Galt P. Barber[9], Laurence R. Meyer[9], Cricket A. Sloan[9], Venkat S. Malladi[9], Melissa S. Cline[9], Katrina Learned[9], Vanessa K. Swing[9], Ann S. Zweig[9], Brooke Rhead[9], Pauline A. Fujita[9], Krishna Roskin[9], Donna Karolchik[9], Robert M. Kuhn[9], David Haussler[9,36].

Data Analysis Center. Ewan Birney[4], Ian Dunham[4], Steven P. Wilder[4], Damian Keefe[4], Daniel Sobral[4], Javier Herrero[4], Kathryn Beal[4], Margus Lukk[4], Alvis Brazma[4], Juan M. Vaquerizas[4], Nicholas M. Luscombe[4], Peter J. Bickel[45], Nathan Boley[45], James B. Brown[45], Qunhua Li[45], Haiyan Huang[45], Mark Gerstein[32,33,34], Lukas Habegger[33], Andrea Sboner[33,34], Joel Rozowsky[33,34], Raymond K. Auerbach[33], Kevin Y. Yip[33,34], Chao Cheng[33,34], Koon-Kiu Yan[33,34], Nitin Bhardwaj[33,34], Jing Wang[33,34], Lucas Lochovsky[33,34], Justin Jee[33,34], Theodore Gibson[33,34], Jing Leng[33,34], Jiang Du[35], Ross C. Hardison[5], Robert S. Harris[5], Giltae Song[5], Webb Miller[5], David Haussler[9,36], Krishna Roskin[9], Bernard Suh[9], Ting Wang[46], Benedict Paten[9], William S. Noble[2,47], Michael M. Hoffman[2], Orion J. Buske[2], Zhiping Weng[48], Xianjun Dong[48], Jie Wang[48], Hualin Xi[49].

University of Albany SUNY Group. Scott A. Tenenbaum[50], Frank Doyle[50], Luiz O. Penalva[51], Sridar Chittur[50].

Boston University Group. Thomas D. Tullius[52], Stephen C. J. Parker[28,52].

University of Chicago, Stanford Group. *University of Chicago*: Kevin P. White[53], Subhradip Karmakar[53], Alec Victorsen[53], Nader Jameel[53], Nick Bild[53], Robert L. Grossman[53]. *Stanford*: Michael Snyder[3], Stephen G. Landt[3], Xinqiong Yang[3], Dorrelyn Patacsil[3], Teri Slifer[3].

University of Massachusetts Medical School Groups. *University of Massachusetts Medical School I*: Job Dekker[44], Bryan R. Lajoie[44], Amartya Sanyal[44]. *University of Massachusetts Medical School II*: Zhiping Weng[48], Troy W. Whitfield[48], Jie Wang[48], Patrick J. Collins[54], Nathan D. Trinklein[54], E. Christopher Partridge[1], Richard M. Myers[1].

Boise State University/University of North Carolina–Chapel Hill Proteomics Group. Morgan C. Giddings[55,56,57], Xian Chen[58], Jainab Khatun[55], Chris Maier[55], Yanbao Yu[57], Harsha Gunawardena[57], Brian Risk[56].

NIH Project Management Group. Elise A. Feingold[58], Rebecca F. Lowdon[58], Laura A. L. Dillon[58], Peter J. Good[58].

Affiliations

1 HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, United States of America,

2 Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America,

3 Department of Genetics, Stanford University School of Medicine, Stanford, California, United States of America,

4 European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridgeshire, United Kingdom,

5 Center for Comparative Genomics and Bioinformatics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania, United States of America,

6 Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America,

7 Howard Hughes Medical Institute and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, United States of America,

8 Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America,

9 Center for Biomolecular Science and Engineering, University of California, Santa Cruz, Santa Cruz, California, United States of America,

10 Biology Division, California Institute of Technology, Pasadena, California, United States of America,

11 Beckman Institute, California Institute of Technology, Pasadena, California, United States of America,

12 Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina, United States of America,

13 Department of Pediatrics, Duke University, Durham, North Carolina, United States of America,

14 Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America,

15 Affymetrix, Santa Clara, California, United States of America,

16 Karolinksa Institutet, Huddinge, Sweden,

17 University of Geneva, Geneva, Switzerland,

18 Bioinformatics and Genomics, Centre de Regulacio Genomica, Barcelona, Spain,

19 Omics Science Center, RIKEN Yokohama Institute, Yokohama, Kanagawa, Japan,

20 Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America,

21 Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland,

22 Genome Institute of Singapore, Singapore,

23 Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, Section of Molecular Genetics and Microbiology, University of Texas at Austin, Austin, Texas, United States of America,

24 Department of Biology, Carolina Center for Genome Sciences, and Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America,

25 Center for Advanced Computing Research, California Institute of Technology, Pasadena, California, United States of America,

26 Department of Pathology, Stanford University School of Medicine, Stanford, California, United States of America,

27 Department of Computer Science, Stanford University, Stanford, California, United States of America,

28 Division of Intramural Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America,

29 National Human Genome Research Institute, Genome Technology Branch, National Institutes of Health, Rockville, Maryland, United States of America,

30 National Human Genome Research Institute, NIH Intramural Sequencing Center, National Institutes of Health, Bethesda, Maryland, United States of America,

31 Vertebrate Genome Analysis, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom,

32 Center for Genome Sciences and Department of Computer Science, Washington University in St. Louis, St. Louis, Missouri, United States of America,

33 Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America,

34 Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, United States of America,

35 Department of Computer Science, Yale University, New Haven, Connecticut, United States of America,

36 Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, California, United States of America,

37 Structural Computational Biology, Centro Nacional de Investigaciones Oncolùgicas, Madrid, Spain,

38 Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut, United States of America,

39 Department of Genetics, Yale University, New Haven, Connecticut, United States of America,

40 Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, United States of America,

41 Department of Pediatrics, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America,

42 Department of Biochemistry and Molecular Biology, Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, United States of America,

43 Genome Center, University of California–Davis, Davis, California, United States of America,

44 Program in Gene Function and Expression, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America,

45 Department of Statistics, University of California at Berkeley, Berkeley, California, United States of America,

46 Department of Genetics, Washington University in St. Louis, St. Louis, Missouri, United States of America,

47 Department of Computer Science and Engineering, University of Washington, Seattle, Washington, United States of America,

48 Program in Bioinformatics and Integrative Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America,

49 Bioinformatics Program, Boston University, Boston, Massachusetts, United States of America,

50 College of Nanoscale Sciences and Engineering, University at Albany–SUNY, Albany, New York, United States of America,

51 Children's Cancer Research Institute, Department of Cellular and Structural Biology, San Antonio, Texas, United States of America,

52 Department of Chemistry and Program in Bioinformatics, Boston University, Boston, Massachusetts, United States of America,

53 Institute for Genomics and Systems Biology, The University of Chicago, Chicago, Illinois, United States of America,

54 SwitchGear Genomics, Menlo Park, California, United States of America,

55 Biomolecular Research Center, Boise State University, Boise, Idaho, United States of America,

56 Department of Microbiology and Immunology, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America,

57 Biochemistry Department, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America,

58 National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America

References

1. Collins FS, Green ED, Guttmacher AE, Guyer MS (2003) A vision for the future of genomics research. Nature 422: 835–847.
2. ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636–640.
3. 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. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.
5. Chiaromonte F, Weber RJ, Roskin KM, Diekhans M, Kent WJ, Haussler D (2003) The share of human genomic DNA under selection estimated from human-mouse genomic alignments. Cold Spring Harb Symp Quant Biol 68: 245–254.
6. Stone EA, Cooper GM, Sidow A (2005) Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. Annu Rev Genomics Hum Genet 6: 143–164.
7. Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH (2009) Local DNA topography correlates with functional noncoding regions of the human genome. Science 324: 389–392.
8. Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, Stamatoyannopoulos JA (2007) Widely distributed noncoding purifying selection in the human genome. Proc Natl Acad Sci U S A 104: 12410–12415.
9. Wold B, Myers RM (2008) Sequence census methods for functional genomics. Nature Meth 5: 19–21.
10. Wang Z, Gerstein M, Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63.
11. Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, et al. (2010) ENCODE whole-genome data in the UCSC Genome Browser. Nucleic Acids Res 38 (Database issue): D620–D625.
12. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, et al. (2010) The UCSC Genome Browser database: update 2010. Nucleic Acids Res 38 (Database issue): D613–D619.
13. Lozzio CB, Lozzio BB (1975) Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. Blood 45: 321–334.
14. Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, et al. (1998) Embryonic stem cell lines derived from human blastocysts. Science 282: 1145–1147.
15. Gey GO, Coffman WD, Kubicek MT (1952) Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. Cancer Res 12: 264–265.
16. Knowles BB, Howe CC, Aden DP (1980) Human hepatocellular carcinoma cell lines secrete the major plasma proteins and hepatitis B surface antigen. Science 209: 497–499.
17. Jaffe EA, Nachman RL, Becker CG, Minick CR (1973) Culture of human endothelial cells derived from umbilical veins: Identification by morphologic and immunologic criteria. J Clin Invest 52: 2745–2756.
18. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931–945.
19. Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, et al. (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol 7 Suppl 1: S2 1–31.
20. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, et al. (2006) GENCODE: producing a reference annotation for ENCODE. Genome Biol 7: S4 1–9.

21. Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M (2006) PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics 22: 1437–1439.
22. Flicek P, Aken BL, Ballester B, Beal K, Bragin E, et al. (2010) Ensembl's 10th year. Nucleic Acids Res 38: D557–D562.
23. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, et al. (2004) Novel RNAs identified from a comprehensive analysis of the transcriptome of human chromosomes 21 and 22. Genome Res 14: 331–342.
24. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods 5: 621–628.
25. Schmid M, Jensen TH (2010) Nuclear quality control of RNA polymerase II transcripts. J Wiley Interdisciplinary Review.
26. Bachellerie JP, Cavaille J, Huttenhofer A (2002) The expanding snoRNA world. Biochimie 84: 775–790.
27. Kapranov P, Willingham AT, Gingeras TR (2007) Genome-wide transcription and the implications for genomic organization. Nat Rev Genet 8: 413–423.
28. Fullwood MJ, Wei CL, Liu ET, Ruan Y (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res 19: 521–532.
29. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa 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 U S A 100: 15776–15781.
30. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38: 626–635.
31. Valen E, Pascarella G, Chalk A, Maeda N, Kojima M, et al. (2009) Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. Genome Res 19: 255–265.
32. Frohman MA, Dush MK, Martin GR (1988) Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc Natl Acad Sci U S A 85: 8998–9002.
33. Giddings MC, Shah AA, Gesteland R, Moore B (2003) Genome-based peptide fingerprint scanning. Proc Natl Acad Sci U S A 100: 20–25.
34. Merrihew GE, Davis C, Ewing B, Williams G, Käll L, et al. (2008) Use of shotgun proteomics for the identification, confirmation, and correction of C elegans gene annotations. Genome Res 18: 1660–1669.
35. Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet 7: 29–59.
36. Wu C (1980) The 5′ ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature 286: 854–860.
37. Keene MA, Corces V, Lowenhaupt K, Elgin SC (1981) DNase I hypersensitive sites in Drosophila chromatin occur at the 5′ ends of regions of transcription. Proc Natl Acad Sci U S A 78: 143–146.
38. McGhee JD, Wood WI, Dolan M, Engel JD, Felsenfeld G (1981) A 200 base pair region at the 5′ end of the chicken adult beta-globin gene is accessible to nuclease digestion. Cell 27: 45–55.
39. Gross DS, Garrard WT (1988) Nuclease hypersensitive sites in chromatin. Annu Rev Biochem 57: 159–197.
40. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 17: 877–885.
41. Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, et al. (2009) Mapping accessible chromatin regions using Sono-Seq. Proc Natl Acad Sci U S A 106: 14926–14931.
42. Kouzarides T (2007) Chromatin modifications and their function. Cell 128: 693–705.
43. Bernstein BE, Meissner A, Lander ES (2007) The mammalian epigenome. Cell 128: 669–681.
44. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching C, et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39: 311–318.
45. Liang G, Lin JC, Wei V, Yoo C, Cheng J, et al. (2004) Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. Proc Natl Acad Sci U S A 11: 7357–7362.
46. Schneider R, Bannister AJ, Myers FA, Thorne AW, Crane-Robinson C, Kouzarides T (2004) Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. Nat Cell Biol 6: 73–77.
47. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, et al. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 28: 169–181.
48. Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, et al. (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature 18: 349–353.
49. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, et al. (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 21: 301–313.
50. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315–326.
51. Strahl BD, Allis CD (2000) The language of covalent histone modifications. Nature 6: 41–45.
52. Sabo PJ, Hawrylycz M, Wallace JC, Humbert R, Yu M, et al. (2004) Discovery of functional noncoding elements by digital analysis of chromatin structure. Proc Natl Acad Sci U S A 101: 16837–16842.
53. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, et al. (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell 25: 311–322.
54. Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, et al. (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat Methods 3: 511–518.
55. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, et al. (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods 6: 283–289.
56. Sekimata M, Pérez-Melgosa M, Miller SA, Weinmann AS, Sabo PJ, et al. (2009) CCCTC-binding factor and the transcription factor T-bet orchestrate T helper 1 cell-specific structure and function at the interferon-gamma locus. Immunity 31: 551–564.
57. Giresi PG, Lieb JD (2009) Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). Methods 48: 233–239.
58. Gaulton KJ, Nammo T, Pasquail L, Simon JM, Giresi PG, et al. (2010) A map of open chromatin in human pancreatic islets. Nat Genet 42: 255–259.
59. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823–837.
60. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497–1502.
61. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553–560.
62. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4: 651–657.
63. Poser I, Sarov M, Hutchins JR, Hériché JK, Toyoda Y, et al. (2008) BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. Nat Methods 5: 409–415.
64. Hua S, Kittler R, White KP (2009) Genomic antagonism between retinoic acid and estrogen signaling in breast cancer. Cell 137: 1259–1271.
65. Raha D, Hong M, Snyder M (2010) ChIP-seq: a method for global identification of regulatory elements in the genome. Curr Protoc Mol Biol Chapter 21: Unit 2119 1–14.
66. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, et al. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 128: 1231–1245.
67. Jaenisch R (1997) DNA methylation and imprinting: Why bother? Trends Genet 13: 323–329.
68. Bird A (2002) DNA methylation patterns and epigenetic memory. Genes & Dev 16: 6–21.
69. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, et al. (2010) Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell 17: 510–522.
70. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger SJ, et al. (2010) Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res 20: 440–446.
71. Laurent L, Wong E, Li G, Tsirigos A, Ong CT, et al. (2010) Dynamic changes in the human methylome during differentiation. Genome Res 20: 320–331.
72. Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, et al. (2010) Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res 20: 434–439.
73. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, et al. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454: 766–770.
74. Brunner AL, Johnson DS, Kim SW, Valouev A, Reddy TE, et al. (2009) Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. Genome Res 19: 1044–1056.
75. Galas DJ, Schmitz A (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res 5: 3157–3170.
76. Strauss EC, Orkin SH (1992) In vivo protein-DNA interactions at hypersensitive site 3 of the human beta-globin locus control region. Proc Natl Acad Sci U S A 89: 5809–5813.
77. Boyle AP, Song L, Lee BK, London D, Keefe D, et al. (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res, in press.
78. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res, in press.
79. Lu Y, Dimasi DP, Hysi PG, Hewitt AW, Burdon KP, et al. (2010) Common genetic variants near the Brittle Cornea Syndrome locus ZNF469 influence the blinding disease risk factor central corneal thickness. PLoS Genet 6: e1000947. doi:10.1371/journal.pgen.1000947.
80. McDaniell R, Lee BK, Song L, Liu Z, Boyle AP, et al. (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. Science 328: 235–239.

81. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, et al. (2010) Variation in transcription factor binding among humans. Science 328: 232–235.
82. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444–454.
83. Korbel JO, Urban AE, Grubert F, Du J, Royce TE, et al. (2007) Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. Proc Natl Acad Sci U S A 104: 10110–10115.
84. Conrad DF, Bird C, Blackburne B, Lindsay S, Mamanova L, et al. (2010) Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. Nat Genet 42: 385–391.
85. Miele A, Dekker J (2008) Long-range chromosomal interactions and gene regulation. Mol Biosyst 4: 1046–1057.
86. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, et al. (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 16: 1299–1309.
87. Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. Science 15: 1306–1311.
88. Lajoie BR, van Berkum NL, Sanyal A, Dekker J (2009) My5C: web tools for chromosome conformation capture studies. Nat Methods 6: 690–691.
89. Baroni TE, Chittur SV, George AD, Tennenbaum SA (2008) Advances in RIP-chip analysis: RNA-binding protein immunoprecipitation-microarray profiling. Methods Mol Biol 419: 93–108.
90. Keene JD, Komisarow JM, Friedersdorf MB (2006) RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nat Protoc 1: 302–307.
91. Tenenbaum SA, et al. (2000) Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. Proc Natl Acad Sci U S A 97: 14085–14090.
92. Tenenbaum SA, Lager PJ, Carson CC, Keene JD (2002) Ribonomics: identifying mRNA subsets in mRNP complexes using antibodies to RNA-binding proteins and genomic arrays. Methods 26: 191–198.
93. Trinklein ND, Karaöz U, Wu J, Halees A, Force Aldred S, et al. (2007) Integrated analysis of experimental data sets reveals many novel promoters in 1% of the human genome. Genome Res 17: 720–731.
94. Lin JM, Collins PJ, Trinklein ND, Fu Y, Xi H, et al. (2007) Transcription factor binding and histone modifications in human bidirectional promoters. Genome Res 17: 818–827.
95. Collins PJ, Kobayashi Y, Nguyen L, Trinklein ND, Myers RM (2007) The ets-related transcription factor GABP directs bidirectional transcription. PLoS Genet 3(11): e208. doi:10.1371/journal.pgen.0030208.
96. Petrykowska HM, Vockley CM, Elnitski L (2008) Detection and characterization of silencers and enhancer-blockers in the greater CFTR locus. Genome Research 18: 1238–1246.
97. Landolin JM, Johnson DS, Trinklein ND, Aldred SF, Medina C, et al. (2010) Sequence features that drive human promoter function and tissue specificity. Genome Res 20: 890–898.
98. Khatun J, Hamlett E, Giddings MC (2008) Incorporating sequence information into the scoring function: a hidden Markov model for improved peptide identification. Bioinformatics 24: 674–681.
99. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E (2008a) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res 8: 1814–1828.
100. Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P (2008b) Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res 18: 1829–1843.
101. Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15: 901–913.
102. Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S (2007) Analysis of sequence conservation at nucleotide resolution. PLoS Comput Biol 3: e254. doi:10.1371/journal.pcbi.0030254.
103. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20: 110–121.
104. Wu JQ, Habegger L, Noisa P, Szekely A, Qiu C, et al. (2010) Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. Proc Natl Acad Sci U S A 107: 5254–5259.
105. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, et al. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 27: 66–75.
106. Toronto International Data Release Workshop Authors, Birney E, Hudson TJ, Green ED, Gunter C, et al. (2009) Prepublication data sharing. Nature 461: 168–170.
107. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.
108. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362–9367.
109. Wokolorczyk D, Gliniewicz B, Sikorski A, Zlowocka E, Masojc B, et al. (2008) A range of cancers is associated with the rs6983267 marker on chromosome 8. Cancer Res 68: 9982–9986.
110. Curtin K, Lin WY, George R, Katory M, Shorto J, et al. (2009) Meta association of colorectal cancer confirms risk alleles at 8q24 and 18q21. Cancer Epidemiol Biomarkers Prev 18: 616–621.
111. Al Olama AA, Kote-Jarai Z, Giles GG, Guy M, Morrison J, et al. (2009) Multiple loci on 8q24 associated with prostate cancer susceptibility. Nat Genet 41: 1058–1060.
112. Jia L, Landan G, Pomerantz M, Jaschek R, Herman P, et al. (2009) Functional enhancers at the gene-poor 8q24 cancer-linked locus. PLoS Gene 5: e1000597. doi:10.1371/journal.pgen.1000597.
113. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, et al. (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet 41: 882–884.
114. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, et al. (2009) The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet 41: 885–890.
115. Wright JB, Brown SJ, Cole MD (2010) Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. Mol Cell Biol 30: 1411–1420.
116. Kidd JM, Sampas N, Antonacci F, Graves T, Fulton R, et al. (2010) Characterization of missing human genome sequences and copy-number polymorphic insertions. Nat Methods 7: 365–371.
117. Li Q, Brown JB, Huang H, Bickel P (2011) Measuring reproducibility of high-throughput experiments. Annals of Applied Statistics, in press.
118. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10: 252–263.
119. Nelson SB, Sugino K, Hempel CM (2006) The problem of neuronal cell types: a physiological genomics approach. Trends Neurosci 29: 339–345.
120. Reisman D, Bálint é, Loging WT, Rotter V, Almon E (1996) A novel transcript encoded within the 10-kb first intron of the human p53 tumor suppressor gene (D17S2179E) is induced during differentiation of myeloid leukemia cells. Genomics 38: 364–370.
121. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet 36: 40–45.