Almost exactly 10 years ago, a “Workshop on the Comprehensive Extraction of Biological Information from Genomic Sequence” endorsed the formation of a public consortium to undertake comprehensive annotation of all functional elements encoded in the human genome, a logical encore to the production of the genome sequence itself. At that time, in spite of general optimism for the cause, few things were clear. The task lacked precedent, obscuring its true scope; the requisite technologies were either nascent or, at the time of the workshop, not yet imagined; and, on the heels of the mouse genome sequence, the very role of experimental approaches was challenged by the burgeoning promise of comparative genomics. What was not in doubt was the commitment of NHGRI to build on the success of the Human Genome Project, which had yet to announce its finished sequence.

In this environment, The ENCODE Project Consortium (ENCODE) took form. A four-year pilot phase initiated in 2003 focused on a carefully selected 1% of the human genome and was oriented chiefly toward the deployment at scale and assessment of experimental and computational genomic technologies for localization of functional elements. In 2007, the pilot project was succeeded by the deployment of ENCODE annotation efforts to the entire genomic sequence. In a happy and fateful coincidence, the ENCODE scale-up commenced contemporaneously with the introduction of massively parallel “next-generation” sequencing technologies—a development that was rapidly exploited by ENCODE groups to replace wholesale the assays that had been utilized during the pilot phase. Indeed, ENCODE groups played leading roles in the genesis and deployment of many staple genomic assays in wide use today, including the first ChIP-plus-sequencing assays (ChIP-seq) for transcription factors (Johnson et al. 2007; Robertson et al. 2007) and histone modifications (Barski et al. 2007; Mikkelsen et al. 2007), as well as pioneering RNA sequencing assays (RNA-seq) (Mortazavi et al. 2008), among others.

The ENCODE production phase has produced thousands of genome-wide data sets during the past five years, yielding deep insights into genome function, and ENCODE is now poised for a further multi-year expansion phase. As such, the present juncture provides a useful vantage from which to reflect on the ENCODE’s accomplishments, challenges, and prospects. Here, I first discuss how ENCODE has influenced our conception of genome structure and content, and the utility of function-driven versus purely sequence-based approaches to genome annotation. Second, I consider major trends that are shaping both the nature of ENCODE data and how those data are conceptualized and used. Finally, I discuss key challenges confronting the next phase of the ENCODE endeavor.

Reading the living genome

Functional elements, then and now

Although the ENCODE project formally originated in the post-genome era, its intellectual origins lie some 40 years earlier with the concept that genomes contain discrete, linearly ordered units that can be connected with specific functional features or processes (Jacob and Monod 1961). A cornerstone of ENCODE has been the use of biochemical signatures to identify functional elements specified by the genomic sequence. In part, this represents a departure from the widely accepted reductionist approach to genome function, in which iterative dissection by truncation or editing of larger sequences that encompass a given functional activity was coupled to an experimental read-out of that activity. The reductionist approach provided a powerful experimental paradigm and was widely applied to define and understand the signals that direct transcription initiation, splicing, and other basic processes, and to expose the transcription factor binding elements that comprise the sequence “atoms” of gene regulation.

The biochemical signature strategy, which developed in parallel with reductionism, was motivated by the recognition of common biochemical or biophysical events that invariably attended certain types of noncoding functional elements. This strategy found its first expression in the discovery that active promoters were marked by alterations in chromatin structure that gave rise to nucleosome hypersensitivity of the underlying DNA (Wu et al. 1979; Wu 1980). This signature was subsequently sought over entire genomic loci (Stalder et al. 1980) and resulted in definition of the first cellular enhancers (Banerji et al. 1983) and other types of transcriptional control elements (Forrester et al. 1986; Grovers et al. 1987; Chung et al. 1993). Reductionism was, in turn, applied to the biochemically defined elements, revealing them to be densely populated by recognition sequences for DNA-binding proteins (Emerson et al. 1985; Strauss and Orkin 1992), motivating, in turn, the development of site-specific factor occupancy assays such as ChIP (Gilmour and Lis 1984; Solomon and Varshavsky 1985). Subsequently, the recognition that histone modification patterns could suggest transcription factor occupancy patterns (Lee et al. 1993) and functional characteristics of adjacent regul-
What does our genome encode?

The dual concept of the gene both as the agent of heredity and as a physical, information-laden entity embodied in a specific DNA sequence has dominated modern biology. Great emphasis has been placed on the accurate and comprehensive annotation of genes in the human genome and across the spectrum of sequenced organisms. Over the last 10 years, ENCODE data have engendered numerous fundamental observations concerning the organization of transcription that have collectively provided deep insights into the genic landscape. The past five years in particular have witnessed a striking convergence of our views connecting chromatin and transcription, spurred by observations grounded in ENCODE data.

Transcription originating from enhancer elements was first described over 20 years ago (Tuan et al. 1992) and has recently reemerged through analysis of deep RNA-seq data (Kim et al. 2012). This phenomenon has now been extensively documented by ENCODE (Djebali et al. 2012a) and is detected predominantly at distal DNase I hypersensitive sites that are flanked by H3K4me1, H3K27ac, and H3K9ac histone modifications. Unlike promoters, the enhancer-originated transcripts derive overwhelmingly from nuclear nonpolyadenylated RNA and are thus missing the large fraction of polyadenylated transcripts characteristic of canonical promoters. The rate of enhancer-originated transcription varies widely, and generally occurs at a substantially lower level than promoters.

The association of trimethylation at histone H3 lysine 4 (H3K4me3) with transcription initiation at human genes is well described (Wang et al. 2008; Ernst et al. 2011) and is a direct outgrowth of studies in yeast (Bernstein et al. 2002; Santos-Rosa et al. 2002). More striking and unexpected was the discovery that...
patterns of histone modification (H3K36me3) and nucleosome location preference within gene bodies reflect organizational features of mature transcript structure such as exons, as well as their splicing frequency (Kolaisnka-Zwierz et al. 2009; Tilgner et al. 2009). However, both the mechanisms giving rise to these phenomena and their implications for global genome function remain largely obscure. It is likely, however, that many more subtle connections between chromatin modification and transcript structure lie waiting to be uncovered in ENCODE data, some of which may be brought to light with the increasing cost effectiveness of deeper sequencing that will, in turn, enable finer parsing of the chromatin landscape.

How the nuclear machinery executes a high-precision operation such as splicing over genomic distances that may exceed 1 Mb is currently unknown. The most straightforward explanation is that, analogous to enhancers and their target promoters, these transcript components are physically approximated to one another through direct chromatin interactions. Such hypotheses are now directly testable by cross-analysis of long-range splicing data with ENCODE 5C (Sanyal et al. 2012) and ChIA-PET results (Li et al. 2012).

Viewed collectively, ENCODE data are increasingly pointing to the conclusion that chromatin and transcription are not discrete genomic forces that collide in the context of gene expression. Rather, they represent a continuum of activities, from the infrequent generation of transcripts at distal regulatory DNA, to regions of high transcriptional output that are marked by pervasive alterations in chromatin state. It is certain that many additional features of the transcription–chromatin connection remain to be uncovered within extant ENCODE data. Going forward, deeper probing of both the RNA and chromatin compartments through advancing sequencing throughput will perhaps bring these features to light more quickly.

Regulatory DNA: More than meets the eye

It is still widely believed that functional elements, from exons to regulatory DNA, are relatively rare features of the genomic landscape. In the case of regulatory DNA, this is certainly true within the context of an individual cell type, where DNase I hypersensitive sites and associated transcription factor occupancy sites mapped by ChIP-seq encompass on the order of 1%–2% of the genome—a compartment roughly the size of the exome. However, because the majority of regulatory DNA regions are highly cell type-selective (The ENCODE Project Consortium 2012; Thurman et al. 2012), the genomic landscape rapidly becomes crowded with regulatory DNA as the number of cell types and states assayed increases. Even after assaying more than 120 distinct cell types, this trend shows little evidence of saturation (The ENCODE Project Consortium 2012). It is thus not unreasonable to expect that 40% and perhaps more of the genome sequence encodes regulatory information—a number that would have been considered heretical at the outset of the ENCODE project. It is important to recognize, however, that this figure encompasses regulatory regions wherein only a subset of the individual nucleotides are under strong evolutionary constraint, such as those at critical contact positions for transcription factor recognition (Neph et al. 2012b).

It is also widely assumed that roughly half of the human genome sequence has been laid waste by transposable elements and other classes of repetitive sequences, which have repeatedly and haphazardly pummeled the genome at various evolutionary intervals. These regions were all but invisible during the ENCODE pilot phase, where they were intentionally masked from microarray designs. But this situation changed dramatically with sequence tag-based assays: Even modest read lengths (36–50 bp) have the potential to align uniquely with over 85% of the genome sequence, and thus to annotate a majority of transposable elements. In marked contrast to the prevailing wisdom, ENCODE chromatin and transcription studies now suggest that a large number of transposable elements encode highly cell type-selective regulatory DNA that controls not only their own cell-selective transcription, but also those of neighboring genes (Djebali et al. 2012a; Thurman et al. 2012). Far from an evolutionary dustbin, transposable elements appear to be active and lively members of the genomic regulatory community, deserving of the same level of scrutiny applied to other genic or regulatory features.

Leaving the flat genome behind

Gene regulation is fundamentally a three-dimensional (3D) process involving dynamic interactions between genomic DNA and the nuclear protein machinery. And yet, common conceptions of regulatory genomic processes are typically unidimensional, playing out over linear genome distance versus physical nuclear distance. ENCODE has played a key role in shifting this paradigm by providing key insights into the topology of gene regulation at two levels—nuclear structure and organization, and physical connectivity among cis-regulatory elements.

From our present vantage, it may seem remarkable that the discovery of the connection between 3D nuclear chromatin architecture and mammalian gene regulation (Weintraub and Groudine 1976) predated many one-dimensional representations that have populated the literature for the past three decades. Enabled by serial innovations in the quantitative analysis of chromatin interactions by Dekker and colleagues (Dekker et al. 2002; Dostie et al. 2006; Lieberman-Aiden et al. 2009; Nora et al. 2012), the visionary conception of Weintraub and Groudine is systematically taking form within the context of ENCODE. HiC data (Lieberman-Aiden et al. 2009) now provide global contact maps of nuclear chromatin that are sufficiently detailed as to enable reconstruction of the folding pattern of chromosomes within the confines of the nucleus, and to define major chromatin compartments. A key challenge is resolution, which increases only slowly with multiplicative increases in sequencing depth. However, recent increases in sequencing throughput have enabled deeper sampling, shedding further light on large-scale chromatin interactions and chromosomal domain architectures (Dixon et al. 2012; Nora et al. 2012).

The discovery of long-range cis-regulatory elements such as the immunoglobulin enhancer (Banerji et al. 1983) and the beta-globin Locus Control Region (Forrester et al. 1986; Grosveld et al. 1987) immediately raised the question of how such distal regulatory regions communicate with their target genes(s) and, more broadly, how genes and regulatory DNA are “wired” along a chromosome. Specific physical interactions appear to be a general property of long-range regulatory control, and are directly assayable with 3C (Dekker et al. 2002) or, at many elements in parallel, by 5C (Dostie et al. 2006). Systematic application of 5C to assay all mutual chromatin interactions over The ENCODE Pilot Project regions comprising 1% of the genome has now enabled a comprehensive synthesis of interactions between promoters and distal elements including likely enhancers and CTCF-occupied sites (Sanyal et al. 2012). Composite interaction technologies developed within ENCODE such as ChIA-PET—essentially a combination of ChIP-seq with a chromosome conformation capture assay—are further illuminating the 3D connectivity of human
genes with one another and with their respective controlling elements (Li et al. 2012).

Together, the chromosome conformation capture-based approaches have probed local, domain, and global level interactions across the mammalian genome (de Wit and de Laat 2012). Although great progress in mapping genome connectivity has been achieved in a few short years, significant challenges remain, both technical and conceptual. The resolution of chromatin interaction assays is generally limited by restriction fragments, and these assays have a “blind” spot around anchor regions or around highly interacting elements such as promoters, where nonspecific local interactions may obscure more specific connections.

One of the greatest technical challenges facing ENCODE is to transform its linear genomic signals into nuclear space, without sacrificing resolution. Such a transformation would dramatically close the gap between cis-regulatory architecture and nuclear architecture, bringing us full circle in the journey begun by the pioneering experiments of Weintraub and Groudine more than three decades ago.

From elements to networks

Transcription factors interact with one another at three basic levels: direct protein–protein interactions; cooperative interactions engendered by binding within the same cis-regulatory element; and cross-regulatory interactions resulting from the binding of one transcription factor within the regulatory DNA regions controlling another factor. Extended across all transcription factors active within a given cell type, the last of these creates a transcription factor regulatory network that functions as a coherent system to process complex biological signals and confer robustness (Neph et al. 2012a).

Transcription factor regulatory networks can now be mapped systematically using two types of ENCODE data—ChIP-seq for individual transcription factors (Gerstein et al. 2012) and genomic DNase I footprinting (Neph et al. 2012a,b). The resulting networks can be rendered either as a regulatory “cloud” or as a hierarchy of interacting factors (Gerstein et al. 2012). However, for most users of ENCODE data, the greatest interest will lie in specific subnetworks that comprise the wiring of small cohorts of transcription factors, such as those involved in pluripotency or hematopoietic differentiation (Neph et al. 2012a). Deepening of ENCODE annotations through the survey of increasing numbers of transcription factors by ChIP-seq, and increasing numbers of DNase I footprints by deeper sequencing and survey of additional cell types, will further enrich our understanding of the human transcription factor network and how it feeds back to the level of individual regulatory DNA regions—and ultimately to other chromatin features and transcript production.

Decoding disease

The ENCODE production phase was initiated at the height of excitement over genome-wide association studies, nearly a thousand of which have since been performed. At that time, the prospects for convergence between the two initiatives seemed limited at best. Five years on, building on initial observations (Gaulton et al. 2010; Ernst et al. 2011), it is now apparent that a significant proportion of strongly disease- and trait-associated variants emerging from genome-wide association studies (GWASs) localize within regulatory DNA marked by DNase I hypersensitive sites and selected TFs (The ENCODE Project Consortium 2012; Maurano et al. 2012a; Schaub et al. 2012). Beyond simple enrichment, analysis of an expanded range of cell and tissue types reveals systematic and deep connections between the tissue and developmental stage selectivity with which disease- and trait-associated variants localize within regulatory DNA; the transcription factor recognition sequences perturbed by these variants; and the networks formed by these transcription factors (Maurano et al. 2012a). The ability to connect distal DHSs systematically with their cognate genes (Thurman et al. 2012) has now revealed many links between variants in regulatory DNA and distant genes that plausibly explain the disease associations (Maurano et al. 2012a).

It is also clear that the modest (but highly significant) overall degree of enrichment of disease- and trait-associated variants within regulatory DNA is an inappropriate measure that incorporates both noise within the GWAS data and the heterogeneous mix of cell types examined, many of which are peripheral to certain traits. In contrast, striking cell-selective enrichment of GWAS variants may be observed in pathogenic cell types, for example, the enrichment of variants associated with Crohn’s disease in DHSs from Th1 T cells (The ENCODE Project Consortium 2012) or, even more prominently, in Th17 T cells, which play a leading role in Crohn’s pathogenesis (Maurano et al. 2012a). Significantly, strong cell-selective enrichments are observed for hundreds of variants that fall below the canonical genome-wide significance threshold ($P < 10^{-8}$), suggesting that GWAS signals may encompass the collective quantitative contributions of large numbers of regulatory variants. Of high priority is determining which specific variants within regulatory DNA functionally impact DNA–protein interactions, local chromatin architecture, and the regulation of target genes (Maurano et al. 2012a). Going forward, ENCODE is well-positioned to contribute substantially to this effort. However, to achieve the highest utility for analysis of disease studies, three challenges confront ENCODE. First, care and coordination must be exercised in selecting cell and tissue types, ideally with close input from the disease communities. Second, selection of transcription factors should be well-matched to known aspects of disease physiology or that of pathogenic cell types. Finally, continuously updated maps of connections between distal regulatory DNA and its target gene(s) must be made available in a format that facilitates integration with GWAS variants.

Another disease area in which ENCODE is poised to yield important insights is cancer. Approximately 40 cancer lines of varying origin have been studied using one or more ENCODE methodologies. Three findings stand out: first, that cancer lines harbor a large number of regulatory DNA regions that are not seen in normal cells (Song et al. 2011; Akhtar-Zaidi et al. 2012; Thurman et al. 2012; Vernot et al. 2012). Second, somatic variation in regulatory DNA of cancer cell lines is unequally distributed, with certain neoplasms having significantly higher rates of somatic regulatory variation (The ENCODE Project Consortium 2012). Additionally, the regulatory DNA of immortal malignant cells (and ES cells) harbors increased germline mutation rates (Thurman et al. 2012; Vernot et al. 2012). How or whether these two processes are connected is unknown. Third, the occupancy landscape of CTCF—and possibly other TFs—differs substantially between normal and immortal cells, a proportion of which is linked to DNA methylation patterns (H Wang et al. 2012). In general, many key observations have been enabled by the concurrent availability of cancer genome sequencing data from The Cancer Genome Atlas (TCGA). To develop these observations further, it would seem logical for ENCODE to align future cancer cell type selections as closely as possible with TCGA.
Trends shaping ENCODE data and their interpretation

From regions to bases

At a mechanistic level, most genomic processes operate with nucleotide precision. Currently, however, most ENCODE annotations define regions of tens to hundreds of bases. Closing this resolution gap will be a major challenge going forward. Sites of RNA transcription initiation and termination can generally be mapped at nucleotide level using current approaches. Mapping of DNase I hypersensitivity peaks is giving way to genomic DNase I footprinting (Neph et al. 2012b), which provides nucleotide resolution mapping of protein occupancy sites. However, ChIP-based approaches still require inference. For example, conventional transcription factor ChIP-seq data can only infer that a peak or region contains the protein of interest, with the zone of inference typically spanning 200–300 bp. In cases in which a recognition sequence for the cognate factor is extant, it tends to underlie the peak signal. However, roughly half of ENCODE ChIP-seq peaks lack a cognate motif, and in the case of certain individual factors, the proportion of motif-less peaks may exceed 90% (J. Wang et al. 2012b). Most of these cases are likely due to indirect occupancy through protein–protein versus protein–DNA interactions. The recently described ChIP-exo approach (Rhee and Pugh 2011) has the potential to increase substantially the resolution with which binding sites can be localized by ChIP-seq. However, its sample requirements are high, and it is unclear whether it can be applied to most TFs with the same success as seen with the high-occupancy factor CTCF, for which a very high proportion of the occupancy sites harbor clear recognition sequences. ChIP-exo also does not address the direct versus indirect occupancy dilemma.

The case of histone modifications and variants is more complicated. Because modifications or variants are typically distributed across multiple sequential nucleosomes, resolution to nucleosome level is probably sufficient for most needs. This should be straightforward for focally distributed modifications that typically span a small number of nucleosomes (e.g., H3K4me3, H3K27ac, H3K9ac) and lie immediately adjacent to regulatory DNA regions such as promoters and enhancers. Nucleosomes are well-positioned in these regions (Fu et al. 2008), providing a good substrate for single-nucleosome resolution. In more distal regions, where positioning breaks down, it may not be possible to achieve this resolution, nor is it necessarily required because many of the marks found away from active regulatory DNA are widely distributed.

One solution to the resolution dilemma for both TF and histone modification ChIP-seq is coupling them with chromatin accessibility. For example, coupling TF ChIP-seq and genomic footprinting data from the same cell type enables discrimination of direct versus indirect occupancy sites (Neph et al. 2012b). It should also be possible to couple chromatin accessibility with histone modification data to increase the effective resolution of the latter, at least in the vicinity of regulatory DNA.

Man versus machine

ENCODE data are a natural substrate for pattern discovery via machine learning. In 2007, the application of machine learning techniques to ENCODE Pilot Project data was still nascent, with many approaches such as hidden Markov models adapted from gene-finding applications, or coupled with basic segmentation approaches to integrate across different data types (Thurman et al. 2007). In contrast, nearly every section of the ENCODE integrative paper resulting from the production phase data was driven by machine learning approaches, ranging from advanced segmentation algorithms capable of handling large numbers of diverse data types simultaneously (Ernst and Kellis 2012; Hoffman et al. 2012), to self-organizing maps (A Mortazavi, S Pepke, G Marinov, and B Wold, in prep.), to other hybrid or specialized approaches (The ENCODE Project Consortium 2012). A fundamental result from these approaches was essentially an ab initio demonstration that discrete classes of functional elements are, in fact, encoded by the genome in a manner that matches our long-held perceptions, and that they merely need the right combination of assays to expose them. Different approaches to genomic segmentation essentially converged on the same conclusions concerning specific classes of genomic features including promoters, exons, 3’ ends of genes, CTCF-occupied sites, and even some classes of enhancers. A tacit assumption has been that as more data sets become available, both the power and the resolution of machine learning approaches will increase, somewhat akin to adding more species in a comparative genomic analysis.

A key point remains, however, that the recognition of biological meaning in the output states of machine learning applications is still almost entirely dependent on human-driven syntheses. It is perhaps instructive to observe that the field of gene annotation—the birthplace of genome-directed machine learning—has come to favor the manual curation-driven approach embodied in GENCODE, in which automated algorithms play largely a supporting role. An open question is whether, or to what degree, the assignment of biological meaning to machine-learned states can itself be automated through systematic incorporation of the vast electronic literature.

Whereas ENCODE’s current efforts are focused on the integration of biochemical features, the sheer volume of data now available may enable a renaissance in sequence-driven annotation. It is currently unknown to what degree ENCODE-enabled annotations can be derived directly from the primary genome sequence itself (Noble et al. 2005). In parallel with ENCODE, a number of efforts have focused on de novo annotation of enhancers or transcription factor-bound regions by combining conservation, transcription factor recognition motifs, and gene expression (Pennacchio et al. 2007; Bussen et al. 2012). The extensive availability of ENCODE-type data for both human and mouse (The Mouse ENCODE Consortium 2012) now provides rich training sets to enable a new generation of machine learning applications (Arvey et al. 2012). Particularly promising is the ability not only to discern complex features such as enhancers directly from sequence data, but also to discriminate those active in different cellular environments (Lee et al. 2011). As an excellent and understanding the complex and subtle combinations of sequence features underlying ENCODE annotations, such approaches may extend ENCODE to regions of the genome that specify functional elements active in cell populations that are not feasible—either operationally or economically—to address experimentally.

Seeing the big picture

For most genomic data, the interpretation of biological meaning is closely linked with data visualization. Understanding how different signals are distributed relative to well-studied genes, transcripts, and regulatory DNA regions provides compelling insights into the meaning of different data types, both alone and in combination. In 2007, most ENCODE data tracks could be listed com-
fortably within the center section of a small poster (The ENCODE Project Consortium 2007). By 2009, it had become apparent that simply calling up the data tracks generated by an individual data production center—let alone the entire consortium—was no longer tenable. Data visualization thus emerged as a major challenge, and yet one that attracted relatively few resources—understandably since major efforts were being directed simply to understand how any single ENCODE data type was to be processed properly in the first place.

Data visualization presents challenges at multiple levels, from logical organization to visual representation. Unfortunately, it is difficult to escape the verdict that ENCODE has fallen short of community expectations at both of these levels. ENCODE data sets themselves are currently difficult to locate, and common tasks aiming to represent large numbers of data tracks—such as visualizing the same data type across a range of cell types, or visualizing many different data types such as TFs within the same cell type—frequently overwhelm the current genome browser paradigm. Compounding the problem, ENCODE data no longer exist in isolation. Other large-scale data generation programs such as the Roadmap Epigenomics Project (Roadmap) (Bernstein et al. 2010), which began running in parallel with ENCODE in late 2008, are producing large volumes of many of the same data types studied by ENCODE, such as DNase I hypersensitivity and histone modifications for different cell and tissue types. More often than not, it is desirable to examine the entire range of a given data type, which requires close integration of ENCODE data with Roadmap and other sources.

Some efforts to redress these issues are under way. To address logical organization and integration with Roadmap data, a combined ENCODE–Roadmap genome browser has been implemented (http://www.epigenomebrowser.org). Beginning in 2010, the Roadmap program invested targeted resources in a new generation of visualization tools designed to facilitate the display and manipulation of large numbers of data tracks. As a result, a variety of novel interfaces are now becoming available such as the epigenome visualization hub (Zhou et al. 2011). In addition, new types of data exploration tools are being developed which will enable pattern-based exploration of ENCODE or Roadmap data sets. In many cases, tools have been tied to the UCSC Genome Browser infrastructure. But it is only a matter of time before ENCODE data become formatted for new “lightweight” genome browsers such as JBrowse (Skinner et al. 2009) that permit smooth scaling from bases to chromosomes, and dynamic reorganization and condensation of large numbers of data tracks. Perhaps the greatest visualization challenge is still imminent: As ENCODE transitions from a one- to a three-dimensional view of genome function, completely new tools and modes of representation will be required. Presently, few appreciate the depth of this problem, and thus little systematic effort is being devoted to visualization apart from first-generation utilities developed by the leading data producers (Lajoie et al. 2009).

Signal and noise

Virtually all data resulting from high-throughput assays have a component of noise. The introduction of phred quality scores (Ewing et al. 2008) for Sanger sequencing played a key role in the human genome project, since they enabled both the monitoring of data quality within a single production center and the direct comparison of data generated by independent producers. Maximizing the signal-to-noise ratio of the genomic enrichment assays used by ENCODE is of paramount importance in the context of generating reference data that will be widely used by the community. “Clean,” high-quality data with high signal-to-noise ratios enable both more accurate delineation of individual elements and increased sensitivity (i.e., the recognition of weaker elements that would otherwise be lost in the noise). High-quality data are particularly vital when deep sequencing can yield additional information such as TF footprints (Neph et al. 2012b).

To date, genomic enrichment assays have lacked quality metrics analogous to the phred score. To address this deficit, both ENCODE and Roadmap have active efforts devoted toward development and application of data quality metrics, as well as the formulation of end-to-end experimental standards for ChIP-seq (Landt et al. 2012) and other data types. It is thus anticipated that emerging quality scores for genomic enrichment assays will have a positive impact on the overall quality of ENCODE data, their utility for the community, and their interoperability with data from diverse laboratories.

The evolution of conservation

At the outset of ENCODE in 2003, it was widely assumed that evolutionary conservation would prove to be the ultimate arbiter of functional elements in the human genome sequence—all that was lacking was a sufficiently deep sampling of vertebrate genomes for comparative analysis. Correspondingly, highly conserved noncoding sequences were frequently equated with regulatory DNA. For a variety of reasons, both of these expectations missed the mark widely. Following on studies of transcriptional regulation in the RET locus (Fisher et al. 2006), The ENCODE Pilot Project raised a general alarm: Most elements defined by biochemical signatures lacked strong evolutionary conservation (The ENCODE Project Consortium 2007). Conversely, most highly conserved elements escaped annotation using biochemical or other functional assays (Attanasio et al. 2008; McGaughey et al. 2008; Taher et al. 2011). These initial findings have been considerably amplified by the vast volume of data accumulated during the current production phase (The ENCODE Project Consortium 2012) and by other functional studies (Blow et al. 2012). Using conventional measures, most ENCODE-defined elements are poorly conserved, or negligibly so. The number of highly conserved noncoding sequences with an overlapping biochemical function is considerably higher (now roughly half vs. <10% [proportionally] after the pilot phase)—although this increase is largely a byproduct of the expanded genomic space annotated by ENCODE, without much enrichment for conserved elements. Complicating this picture, many elements lacking strong conventional signatures of purifying evolutionary selection nonetheless appear to be under constraint in human populations (Vernot et al. 2012).

What conclusions should we draw? On a practical level, the ability to measure function at scale has minimized the role of conservation as a discovery tool. But it has also exposed our ignorance concerning the evolutionary forces shaping the genome, particularly in noncoding regions. The fact that per-nucleotide evolutionary conservation, in combination with nucleotide-level DNA accessibility, can accurately trace a protein–DNA binding interface (Neph et al. 2012b) suggests that the operation of purifying selection is vastly more subtle and complexly structured than had been previously assumed. Moreover, nucleotide-level evolutionary conservation is by itself a poor predictor of functional regulatory variation (Maurano et al. 2012b). However, engrained habits of thought are difficult to escape, and highly conserved
noncoding elements are still regularly conflated with regulatory elements (Lowe et al. 2011). Clearly, new models of evolutionary conservation are needed to explain the subtleties of regulatory DNA, and the vast trove of ENCODE data provides an unprecedented opportunity for novel and creative syntheses.

A new entry that promises to reshape the conservation conversation is The Mouse ENCODE Project, from which substantial data are already becoming available (Kim et al. 2012; The Mouse ENCODE Consortium 2012; Shen et al. 2012). These data, which have been generated with the same core experimental pipelines used for human ENCODE, will for the first time enable systematic, genome-wide connections between both sequence and the diverse functional modalities encoded in each species’ genome.

The road ahead
ENCODE has made many seminal contributions and is poised for continued success. However, many challenges remain. Prominent among them are two. First is the question of function itself: How will ENCODE-defined elements be assigned a specific functional property (or properties)? Second, how will ENCODE maximize its utility for the broader scientific community?

Localization versus function
The pre-genome era provided a simple reductionist formula for precisely localizing functional elements and their key internal components: identify, truncate, test—repeat. We learned that operationally defined functional elements such as enhancers and promoters comprise linearly ordered collections of recognition sequences for DNA-binding proteins—the atoms of the regulatory DNA universe. The genomic scale-up of biochemical signature mapping under ENCODE dramatically reshaped this formula, with the ability to delineate likely functional elements greatly outstripping any capacity for defining their functional characteristics through directed experimentation. The result has been a boon for sequence-driven analyses, from regulatory motif derivation to comparative and population genomics.

However, it has also given rise to a broad tendency to think of all elements of a biochemically defined class as having the same functional properties. For example, genomic occupancy by the poly-zinc finger transcriptional regulator CTCF is a prominent feature of experimentally defined enhancer blockers and chromatin boundary elements, as well as bifunctional elements (Gaszner and Felsenfeld 2006). Yet it has now become commonplace to find any CTCF occupancy sites obtained by ChIP-seq referred to as “insulators” without any further specification—and without regard to the well-documented involvement of promoter-bound CTCF in transcriptional control (Klenova et al. 1993). Compounding this complexity, ENCODE has now made available data sets encompassing CTCF occupancy across large numbers of cell types (The ENCODE Project Consortium 2012), revealing substantial diversity in occupancy patterns that reflect important differences in regulation and likely in function (H Wang et al. 2012). Both the sheer number and diversity of these elements argue strongly against ascribing a monolithic functional activity.

A similar situation obtains in the case of enhancers—classically, elements that mediate transcriptional up-regulation, frequently acting at considerable distance from their target gene(s) (Maston et al. 2006). Analysis of ENCODE pilot project data revealed a high ratio of mono- to trimethylated H3K4 at a subset of distal DNase I hypersensitive sites (The ENCODE Project Consortium 2007) and at sites of occupancy by the EP300 (also known as p300) acetyltransferase (Heintzman et al. 2007). In spite of the lack of rigorous functional validation, it has now become de rigueur to refer to any region of the genome that exhibits this combination of modifications as an “enhancer” (Heintzman et al. 2009), and further to characterize “strong” and “weak” enhancers merely on the basis of the intensity of the chromatin modification signal (Ernst et al. 2011), or to designate “poised enhancers” (Creighton et al. 2010) or other subcategorizations based purely on the fine parsing of histone modification patterns (Zentner et al. 2011).

These examples illustrate a natural temptation to equate activity with patterning of epigenomic features. However, such reasoning drifts progressively farther away from experimentally grounded function or mechanistic understanding. The sheer diversity of cross-cell-type regulatory patterning evident in distal regulatory DNA uncovered by ENCODE (Song et al. 2011; Thurman et al. 2012) suggests tremendous heterogeneity and functional diversity. ENCODE is thus in a unique position to promote clearer terminology that separates the identification of functional elements per se from the ascription of specific functional activities using historical experimentally defined categories, and also to dissuade the ascription of very specific functions based on a biochemical signature in place of a deeper mechanistic understanding.

Functional validation: What, how, and how much?
The lack of extensive classical functional validation performed by ENCODE to date is understandable, given the chasm between the number of biochemically defined elements and the throughput of traditional experimental approaches. But what kinds of elements to validate, how to validate them, and how much of each will be considered definitive? Certainly we cannot expect such an effort to be comprehensive; there are too many elements defined in too many cellular contexts ever to validate individually. A logical approach is class-based validation, with the aim to determine, with statistical rigor, how many members of a given class with given biochemically defined features have a specific functional property, in order that a reliable statement may be made. However, it is presently far from clear that we know how properly to categorize the elements we have. Given their diversity, it is likely that a far larger number will need to be examined than would be feasible with conventional methods. In the case of transcriptional enhancer assay by transient transfection, newer high-throughput approaches are emerging (Melnikov et al. 2012; Fatwadhah et al. 2012). However, these impose significant size constraints that restrict their utility. Moreover, the drawbacks of conventional transient assays are well known, most notably the fact that many elements require a chromatin context to function, or a particular primary cellular environment not amenable to transfection.

If significant time and effort is to be invested in high-throughput functional validation, it should be definitive. An emerging alternative that fits this requirement is reverse genetics in an isogenic setting. Once unthinkable for the human genome, knockout of ENCODE-defined regulatory elements is now readily feasible given rapid advances in genome editing technology such as zinc-finger and TAL effector-like nucleases (Doyon et al. 2011; Miller et al. 2011; J Wang et al. 2012a). This technology is currently at a scaling stage (Reyon et al. 2012); given the proper application of resources, thousands of well-designed experiments could reasonably be envisioned over the course of the next phase of ENCODE. Genome editing is well-published in the ENCODE Tier 1 cell type K562 and has the additional advantage of creating a per-
manent reagent (the knockout line) that can be used for more detailed functional characterization by the community.

Completeness: What does it mean?

ENCODE was founded with the ultimate objective of amassing a complete catalog of functional elements encoded by the human genome. Nine years on, we are beginning to appreciate the true scope of this lofty goal. More of the human genome sequence appears to be used for some reproducible, biochemically defined activity than was previously imagined. Contrary to the initial expectations of many, the overwhelming majority of these activities appear to be state-specific—either restricted to specific cell types or lineages, or evolvable in response to a stimulus such as interferon. As such, even if we were in possession of technologies with perfect sensitivity in a given cellular context, the sheer diversity of cell types and states is daunting.

It is becoming increasingly clear that functional annotation of the genome entails understanding not only that a particular stretch of DNA encodes a given type of element active in some cell type, but how that encoding is interpreted in different cellular environments. For example, it is widely acknowledged that the same DNA element may be recognized by different (generally related) transcription factors in different cellular environments, with alternative functional consequences. Additionally, we now know that the biochemical signatures of many ENCODE-defined elements exhibit complex trans-cellular patterns of activity (The ENCODE Project Consortium 2012; Thurman et al. 2012), which may be accompanied by functional behaviors such as an enhancer interacting with different target genes (Sanyal et al. 2012; Thurman et al. 2012). Together, these observations suggest that the genome may, in fact, be extensively multiply encoded—i.e., that the same DNA element gives rise to different activities in different cell types. This possibility challenges our current notions of annotation, which are still rooted in a linear world, and cautions against formulating definitions of completeness based on older models such as the delineation of protein-coding genes.

Maximizing utility for the community

The transition from The ENCODE Pilot Project to the production phase was dominated by technology. Based on many of the trends discussed above, one may predict that the transition from the current production phase to the next will be dominated by utility. Given what we now know about the potential for ENCODE to illuminate not only the genome sequence itself, but also the findings emerging from parallel efforts such as GWAS and TCGA, care must be taken to maximize synergies through careful selection of biological targets and highly coordinated action that maximizes the data generated for each cell or tissue. The high cell requirements entailed by extensive transcription factor ChIP-seq profiling or subcellular RNA fractionation experiments entailed an initial focus on a common set of mainly immortal cells. This has contributed to a perception that ENCODE is largely a cell line-centered endeavor, with limited relevance for many widely studied biological processes. However, overall ENCODE has sampled a vast range of primary cell types—indeed, these outnumber immortalized cell lines nearly 3-to-1 (The ENCODE Project Consortium 2012; Thurman et al. 2012). The potential of ENCODE to contribute to diverse community endeavors is thus now very broad and will be expanded further in the coming production phase as more primary cells enter the experimental pipelines and additional data types such as DNA methylation or maps of RNA-binding proteins become widely available.

ENCODE must recognize and face its awareness problem straight on. ENCODE publications have been cited thousands of times. And yet, broad swathes of the community—even leading-edge laboratories—are unaware of what the project has produced or how to access and interpret the data. A remedy for this situation will not appear spontaneously and will require the intimate involvement of data producers as well as analysts and end users. The only certainty is that if consistent emphasis is not placed on the goal of increasing awareness, and clear milestones defined, little if any progress will be made.

ENCODE is undergoing a transformation from a loosely connected set of annotations to an integrated tool that collectively provides a unique lens through which to view genome function. In this sense, it is gradually transforming from a collection of data into a new kind of tool—almost a type of software that can "operate" on other genomic data types. Indeed, new applications that leverage ENCODE data in this way are already emerging from within the Consortium (Boyle et al. 2012; Ward and Kellis 2012), and one anticipates that many others from diverse community sources are either on the way or will be stimulated as a result of the current suite of ENCODE publications.

Acknowledgments

I thank the anonymous reviewers and the Editor for helpful suggestions; Sam John for insightful discussions and critical reading of the manuscript; and Richard Humbert for help with references. This work was supported in part by NIH grant U54HG004592.

References

Affymetrix/Cold Spring Harbor Laboratory ENCODE Transciptome Project. 2009. Post-transcriptional processing generates a diversity of 5’-modified long and short RNAs. Nature 457: 1028–1032.

Akhtar-Zaidi B, Cowper-Salari R, Conradin O, Salakhova A, Bartels CF, Balasubramanian D, Myeroff L, Lutterbaugh J, Jurra A, Kalady MF, et al. 2012. Epigenomic enhancer profiling defines a signature of colon cancer. Science 336: 736–739.

Arvey A, Aguas P, Noble WS, Leslie C. 2012. Sequence and chromatin determinants of cell-type-specific transcription factor binding. Genome Res (this issue). doi: 10.1101/gr.127712.111.

Attanasio C, Reymond A, Humbert R, Lyle R, Kuehn MS, Nephi S, Sabo PJ, Goldy J, Weaver M, Haydock A, et al. 2008. Assaying the regulatory potential of mammalian conserved non-coding sequences in human cells. Genome Biol 9: R168. doi: 10.1186/gb-2008-9-12-r168.

Banerji J, Olson L, Schaffner W. 1983. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell 33: 729–740.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. Cell 129: 823–837.

Bernstein BE, Humphrey EL, Ezrich RL, Schneider R, Bouman P, Liu JS, Kouzarides T, Schreiber SL. 2006. Methylation of histone H3 lys 4 in coding regions of active genes. Proc Natl Acad Sci 99: 8695–8700.

Bernstein BE, Kamal M, Lindblad-Thor K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ III, Gingeras TR, et al. 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 120: 169–181.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 28: 1045–1048.

Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2012. ChIP-Seq identification of weakly conserved heart enhancers. Nat Genet 42: 2086–2100.

Boyle AP, Hong EL, Hanishan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, et al. 2012. Annotation of
functional variation in personal genomes using RegulomeDB. Genome Res (this issue). doi: 10.1101/gr.137323.112.

Busser BW, Taheaer L, Kim Y, Tansley T, Bloom MJ, Ovcharenko I, Michelson AM. 2012. A machine learning approach for identifying novel cell type-specific transcriptional regulators of myosin 9. PLoS Genet 8:e1002531. doi: 10.1371/journal.pgen.1002531.

Cheng J, Kapranov P, Dekhou J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Heft G, et al. 2005. Transcriptomal maps of 10 human chromosome regions. Science 308:1149–1154.

Chung JH, Whiteley M, Felsenfeld G. 1993. A 5′ element of the chicken β-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. Cell 74:505–514.

Clark AM, Maran SS, Schlesinger RJ, Dinger ME, Bat RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, et al. 2011. The reality of pervasive transcription. PLoS Biol 9:e1000625. doi: 10.1371/journal.pbi.1000625.

Croyton MP, Cheng AW, Webster GG, Koolstra T, Carey BW, Steine JH, Luo M, Fratamico PM, Morris KV, Morillon A, et al. 2011. The ENCODE project consortium. Nature 478:267–279.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. Science 295:1306–1311.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Ernst J, Kellis M. 2012. ChromHMM: Automating chromatin-state discovery and characterization. Genome Res (this issue). doi: 10.1101/gr.135350.111.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485:370–374.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi AM, Tanzer A, Lager A, Ladouceur D, Slattery J, et al. 2012a. Landscape of transcription in human cells. Nature (in press).

Djebali S, Lager A, Kapranov P, Lasa I, Virieux E, Flicek P, Pertea M, Guernec G, Dubin A, Lassmann T, Zakian VA, et al. 2012b. A comprehensive map of the human transcriptome. Science 339:62–67.

Efroni S, Duttagupta R, Cheng J, Dehghani H, Hoeppner DJ, Dash C, Bazett-Jones DP, Le Grice S, McKay RD, Buetow KH, et al. 2008. Global analysis of chromatin state dynamics in nine human cell types. Nat Genet 40:43–49.

Ernst J, Kheradpour P, Mikkelsen TS, Shorson N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473:43–49.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res 8:175–185.

Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science 312:276–279.

Forrester WC, Thompson C, Elder JT, Groudine M. 1986. A developmentally stable chromatin structure in the human β-globin gene cluster. Proc Natl Acad Sci U S A 83:1359–1363.

Frenkel-Morgenstern M, Lacroix V, Ezekurdia I, Levin Y, Gabashvili A, Prilusky J, Del Pozo A, Tress M, Johnson R, Guigo R, et al. 2009. Comparative mapping of 1299–1309.

Girvan NS, Peng R, Cox DW, Sargent MW, Elledge SJ. 2009. Conceptual and technical challenges of high-throughput chromatin immunoprecipitation. Nat Methods 6:90–99.

Goldshmidt A, Nannoni P, Passioli L, Simon JM, Guerin P, Fagni MP, Panhuis TM, Mieczkowski P, Secchi A, Bosco D, et al. 2010. A map of open chromatin in human pancreatic islets. Nature 462:255–259.

Gerstein MB, Kundaje A, Hariharan M, Lindblad-Toh K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, et al. 2012. Architecture of the human regulatory network derived from ENCODE data. Nature (in press).

GMU 2012. Detecting protein–DNA interactions in vivo: Distribution of RNA polymerase on specific bacterial genes. Proc Natl Acad Sci U S A 81:4275–4279.

Grosveld F, van Assendelft GB, Geurts DR, Kollia G. 1987. Position-independent, high-level expression of the human β-globin gene in transgenic mice. Cell 51:975–985.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrett D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human annotation for The ENCODE Project. Genome Res (this issue). doi: 10.1101/gr.135350.111.

Heintzmann RD, Staerk RJ, Hager J, Veltman JA, Bieker LJ, Dekker M, Brown K, et al. 2009. On the spatial organization of active enhancers in the human genome. Nature 458:311–318.

Heintzmann RD, Toth RJ, Hager J, Henni I, Kriete S, Stathopoulu R, Meller J, et al. 2012. Histone modifications and enhancer activity in embryonic stem cells. Nature 495:421–424.

Hoffman MM, Buske OJ, Wang J, Wenz G, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods 9:473–476.

Howlard C, Tanzer A, Chintagunta PK, Derrien T, Walters N, Gonzalez JM, Frankish A, Aken BL, Hourlier T, et al. 2012. Combining RT-PCR-seq and RNA-seq to catalog all genetic elements encoded in the human genome. Genome Res (this issue). doi: 10.1101/gr.134478.111.

Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:315–328.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein–DNA interactions. Science 316:1497–1502.

Kapranov P, Willingham AT, Gingeras TR. 2007. Genome-wide transcription and the implications for genomic organization. Nat Rev Genet 8:413–423.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Lajoie BR, van Berkum NL, Sanyal A, Dekker J. 2009. My5C: Web tools for mapping and analysis of chromatin state transitions in different cell types. Genome Res 19:761–768.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Lajoie BR, van Berkum NL, Sanyal A, Dekker J. 2009. My5C: Web tools for mapping and analysis of chromatin state transitions in different cell types. Genome Res 19:761–768.

Lajoie BR, van Berkum NL, Sanyal A, Dekker J. 2009. My5C: Web tools for mapping and analysis of chromatin state transitions in different cell types. Genome Res 19:761–768.
What does our genome encode?

Mastor GA, Evans SK, Green MR. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 7: 29–59.

Maurano M, Humbert R, Ryne S, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Shafer A, et al. 2012a. Systematic localization of common disease-associated variation in regulatory DNA. *Science* (in press).

Maurano MT, Wang H, Kutyavin T, Stamatoyannopoulos JA. 2012b. Widespread site-dependent buffering of human regulatory polymorphism. *PloS Genet* 8: e1002599. doi: 10.1371/journal.pgen.1002599.

McGaughey DM, Vinton RM, Huynh J, Al-Saif A, Beer MA, McCallion AS, Reyon D, Tsai SQ, Khayter C, Foden JA, Sander JD, Joung JK. 2012. FLASH

Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein–DNA interactions detected at single-nucleotide resolution. *Nature* 504: 5809–5813.

Noble WS, Kuehn S, Thurman R, Yu M, Stamatoyannopoulos J. 2005. Genome-wide profiles of transcription factor regulatory networks. *Genome Res* 15: 2119–2125.

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, Stalder J, Larsen A, Engel JD, Dolan M, Groudine M, Weintraub H. 1980. Tissue-specific DNA cleavages in the globin chromatin domain introduced by DNase I. *Cell* 20: 451–460.

Strauss EC, Orkin SH. 1992. In vivo protein–DNA interactions at hypersensitive site 3 of the human β-globin locus control region. *Proc Natl Acad Sci USA* 89: 5809–5813.

Taher I, McGaughey DM, Maragh S, Aneas I, Bessling SL, Miller W, Nobrega MA, McCallion AS, Ovcharenko I. 2011. Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Res* 21: 1139–1149.

Thurman RE, Day N, Noble WS, Stamatoyannopoulos JA. 2007. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res* 17: 917–927.

Thurman RE, Byrnes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheinin FC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* (in press).

Tilgner H, Nikolau C, Althammer S, Sammelt M, Beato M, Valcarcel J, Guigo R. 2009. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* 16: 996–1001.

Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M, Sandstrom R, et al. 2012. Widespread plasticity in eCTCF occupancy linked to DNA methylation. *Genome Res* (this issue). doi: 10.1101/gr.136127.111.

Wang J, Friedman G, Doyon Y, Wang NS, Li CJ, Miller JC, Hua KL, Yan J, Babiarz JE, Gregory PD, et al. 2012a. Targeted gene addition to a predetermined site in the human genome using a ZFN-based nicking enzyme. *Genome Res* 22: 1316–1326.

Wang J, Zhuang J, Iyer S, Lin XY, Whitfield TW, Greven MC, Pierce RG, Dong X, Kudlacek J, Cheng Y, et al. 2012b. Sequence features and chromatin structure around the genic regions of the human genome but inefficient for lncRNAs. *Genome Res* (this issue). doi: 10.1101/gr.134445.111.

Wang J, Friedman G, Doyon Y, Wang NS, Li CJ, Miller JC, Hua KL, Yan J, Babiarz JE, Gregory PD, et al. 2012a. Targeted gene addition to a predetermined site in the human genome using a ZFN-based nicking enzyme. *Genome Res* 22: 1316–1326.

Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermuller J, Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcarcel J, Guigo R. 2009. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* 16: 996–1001.

Wang J, Friedman G, Doyon Y, Wang NS, Li CJ, Miller JC, Hua KL, Yan J, Babiarz JE, Gregory PD, et al. 2012a. Targeted gene addition to a predetermined site in the human genome using a ZFN-based nicking enzyme. *Genome Res* 22: 1316–1326.

Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermuller J, Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcarcel J, Guigo R. 2009. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* 16: 996–1001.

Wang J, Friedman G, Doyon Y, Wang NS, Li CJ, Miller JC, Hua KL, Yan J, Babiarz JE, Gregory PD, et al. 2012a. Targeted gene addition to a predetermined site in the human genome using a ZFN-based nicking enzyme. *Genome Res* 22: 1316–1326.

Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermuller J, Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcarcel J, Guigo R. 2009. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* 16: 996–1001.

Wang Z, Zhang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Sui K, Roh TY, Peng W, Zhang MQ, et al. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* 40: 897–903.

Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M, Sandstrom R, et al. 2012. Widespread plasticity in eCTCF occupancy linked to DNA methylation. *Genome Res* (this issue). doi: 10.1101/gr.136101.111.

Wang J, Friedman G, Doyon Y, Wang NS, Li CJ, Miller JC, Hua KL, Yan J, Babiarz JE, Gregory PD, et al. 2012a. Targeted gene addition to a predetermined site in the human genome using a ZFN-based nicking enzyme. *Genome Res* 22: 1316–1326.

Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermuller J, Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcarcel J, Guigo R. 2009. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* 16: 996–1001.

Wang J, Friedman G, Doyon Y, Wang NS, Li CJ, Miller JC, Hua KL, Yan J, Babiarz JE, Gregory PD, et al. 2012a. Targeted gene addition to a predetermined site in the human genome using a ZFN-based nicking enzyme. *Genome Res* 22: 1316–1326.