G&I Genomics & Informatics

pISSN 1598-866X eISSN 2234-0742
Genomics Inform 2013;11(2):60-67
http://dx.doi.org/10.5808/GI.2013.11.2.60

REVIEW ARTICLE

Human Transcriptome and Chromatin Modifications: An ENCODE Perspective

Li Shen[1], Inchan Choi[2], Eric J. Nestler[1], Kyoung-Jae Won[2]*

[1]Department of Neuroscience, Mount Sinai School of Medicine, New York, NY 10029, USA,
[2]Department of Genetics, Institute for Diabetes, Obesity and Metabolism,
University of Pennsylvania, Philadelphia, PA 19104, USA

A decade-long project, led by several international research groups, called the Encyclopedia of DNA Elements (ENCODE), recently released an unprecedented amount of data. The ambitious project covers transcriptome, cistrome, epigenome, and interactome data from more than 1,600 sets of experiments in human. To make use of this valuable resource, it is important to understand the information it represents and the techniques that were used to generate these data. In this review, we introduce the data that ENCODE generated, summarize the observations from the data analysis, and revisit a computational approach that ENCODE used to predict gene expression, with a focus on the human transcriptome and its association with chromatin modifications.

Keywords: chromatin modification, ENCODE, GENCODE, transcriptome

Introduction

In September 2012, 30 research papers, including 6 in Nature, were published online (http://www.nature.com/encode/) about genomescale data from a decade-long project, the Encyclopedia of DNA Elements (ENCODE) [1]. Aiming to delineate all functional elements encoded in the human genome [1-4], the ENCODE project examined 1% of the human genome in its pilot phase and scaled up to the whole genome in its second phase. It released data from more than 1,600 sets of experiments from 147 types of tissues [4]. These data include a catalog of human protein-coding and noncoding RNAs as well as protein-DNA interactions, chromatin and DNA accessibility, histone modifications, DNA methylation, and long-range chromosomal interactions. The ENCODE consortium reported that 80.4% of the human genome serves some type of known biochemical function [4]. The unprecedented volume and span of this study will make it an excellent resource for biological analyses. On the other hand, it can be overwhelming for a researcher to access and interpret the data. ENCODE also developed novel (or refined existing) techniques for data generation and computational analyses. These techniques are by no means restricted to ENCODE and can be employed in a wide array of applications. Because of the vast span of ENCODE, we will limit our scope to the human transcriptome and its association with chromatin modifications in this review. We aim to introduce the data that ENCODE generated and the techniques (both experimental and computational) that were used to generate them. We summarize the observations that ENCODE found. We also provide a detailed discussion of a machine learning method that was used to predict gene expression from chromatin modifications with higher accuracy than its predecessors.

Chromatin Modifications Measured by Chromatin Immunoprecipitation Sequencing (ChIP-seq)

DNA sequences are wrapped around octamers of histone proteins to form nucleosomes, the unit of chromatin. The nucleosome core is composed of two copies each of four histone proteins. Each histone has an N-terminal tail that faces outward from the nucleosome and can be chemically modified to influence the accessibility of the chromatin and interactions with other chromatin-binding proteins. These

Received February 14, 2013; Revised March 5, 2013; Accepted March 13, 2013

***Corresponding author:** Tel: +1-215-746-8424, Fax: +1-215-898-5408, E-mail: wonk@mail.med.upenn.edu

Copyright © 2013 by the Korea Genome Organization
It is identical to the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/).

histone modifications are associated with the activation or repression of gene transcription [5] and many other activities of the genome, such as enhancer activities [6] and splicing regulation [7, 8]. There has been great interest in dissecting the interactions between histones and other chromatin modifications and transcriptional regulation in recent years [9]. One of ENCODE's major goals [4] is to use these chromatin modifications to define regulatory elements of the human genome in multiple cell lines and to investigate their interactions with RNA transcription.

Chromatin immunoprecipitation (ChIP) is an experimental technique used to investigate protein-DNA interactions in the cell. ChIP, followed by microarray (ChIP-chip) or high-throughput sequencing (ChIP-seq), is widely used to determine genomewide binding locations of proteins, including transcription factors, covalent modifications of histones, and other chromatin regulatory proteins. ChIP-chip was the chosen technology for the pilot phase of ENCODE to dissect the regulatory regions in the 1% selected portion of the human genome [2, 10, 11]. As next-generation sequencing technology advanced [12], it soon became clear that ChIP-seq was a superior approach [13, 14].

ChIP-seq was the chosen technology for the second phase of the ENCODE project [3, 4, 15]. Although ChIP-seq shows clear advantages over ChIP-chip, it is by no means a perfect technology. In an evaluation study performed by Chen *et al.* [16], multiple factors were found to influence the fidelity of ChIP-seq data. For example, ChIP-seq data were found to be biased toward open chromatin regions, leading to false positives if not corrected; comparison of different algorithms also showed notable variation in sensitivity and specificity. The ENCODE consortium has established a set of guidelines to ensure the quality of ChIP-seq experiments [15]. Because antibodies play a predominant role in the success of ChIP experiments, a significant effort has been made by ENCODE to characterize the specificity and efficiency of a large number of antibodies. The list of the antibodies used and validated by ENCODE can be found at: http://genome.ucsc.edu/ENCODE/antibodies.html. A large-scale assessment of histone modification antibodies (>200) was performed by multiple groups [17] of the ENCODE consortium, with the most up-to-date information available at: http://compbio.med.harvard.edu/antibodies/. In addition, multiple quality metrics have been developed and used by the ENCODE project [15], as listed in Table 1. Having employed these quality metrics, the ENCODE consortium mapped 11 histone modifications plus one histone variant across 46 human cell types, including a complete matrix of the 12 marks across two groups of cell lines, designated as tier 1 and tier 2.

Transcriptome in Human Cells

ENCODE releases a reference gene set (GENCODE) and RNA expression catalogs

The ENCODE project has produced a reference gene set, referred to as GENCODE (http://www.gencodegenes.org) [18]. GENCODE (version 7) identified a comprehensive set of 20,687 protein-coding and 9,640 manually curated long noncoding RNA (lncRNA) loci (representing 15,512 non-coding transcripts/isoforms). Currently, in version 15, it has 20,447 coding genes and 13,249 lncRNA loci (representing

Table 1. Quality metrics of chromatin modification ChIP-seq data employed by ENCODE

Name	Definition	Measures	Requirements
Sequencing depth	The number of uniquely mapped reads	Sufficiency of sequencing	>10 million for sharp peak >20 million for broad peak
NRF	$NRF = \frac{\#\text{uniquely mapped locations}}{\#\text{uniquely mapped reads}}$	PCR bottlenecking	NRF ≥ 0.8
FRiP	$FRiP = \frac{\#\text{mapped reads in peaks}}{\#\text{total mapped reads}}$	Signal-to-noise ratio	FRiP > 1%
CC	$NSC = \frac{CC\ (\text{fragment lenth})}{\min\ (CC)}$ $RSC = \frac{CC\ (\text{fragment lenth}) - \min\ (CC)}{CC\ (\text{phantom peak}) - \min\ (CC)}$	ChIP enrichment	NSC ≥ 1.05 RSC ≥ 0.8
IDR	The posterior probability for a peak being in the irreproducible group	Reproducibility between replicates	IDR < 0.01
Annotation enrichment	Enrichment of chromatin-associated modifications and proteins at functionally annotated features, such as TSS and TES	Known characteristic enrichment	Not specified; based on human experts

ChIP-seq, chromatin immunoprecipitation sequencing; ENCODE, Encyclopedia of DNA Elements; NRF, nonredundancy fraction; PCR, polymerase chain reaction; FRiP, fraction of reads in peaks; CC, cross correlation; NSC, normalized strand cross-correlation; RSC, relative strand cross-correlation; IDR, irreproducible discovery rate; TSS, transcription start site; TES, transcription end site.

Table 2. Comparison of lncRNAs and coding RNAs

Property	lncRNAs	Coding RNAs
Protein coding potential	Very weak	Higher
Structure	Majority has 2 exons (42%) Longer exons and introns	6% has 2 exons, 75% has at least two different/dominant major isoform
Evolutionary conservation	Weaker 30% primate specific	Stronger
Chromatin marks	Active histone marks at TSS Slight excess level of silencing (H3K27me3) and activating (H3K36me3) marks	Active histone marks at TSS
Class	Predominantly in nucleus Significantly more enriched in chromatin than mRNAs	Predominantly in cytosol

LncRNA, long noncoding RNA; TSS, transcriptional start site.

22,531 noncoding transcripts/isoforms). For this reference gene set, GENCODE used manual gene annotation from the Human and Vertebrate Analysis and Annotation (HAVANA) group (http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/) and automatic gene annotation from Ensembl [19]. To construct a comprehensive RNA expression catalog, ENCODE sequenced RNA (using RNA-seq) from 15 different cell lines in multiple subcellular fractions [18]. The expression catalog in multiple cell types provides the transcriptome of coding and noncoding genes, short (<200 bp) and long (>200 bp) in length, polyadenylated or nonpolyadenylated, as well as the compartment in cells where RNAs in each type are populated.

Long noncoding RNAs

LncRNAs are nonprotein-coding transcripts longer than 200 nucleotides. LncRNAs have been known to have similar characteristics as coding RNAs. They are polyadenylated, associated with chromatin signatures, and have multiexonic structure [20, 21]. Some lncRNAs use identical or almost identical transcription initiation complexes [22] and sometimes overlap with protein-coding genes and can be transcribed from either strand [22, 23].

GENECODE interrogated the properties of 15,512 lncRNAs. Table 2 summarizes the comparison between lncRNAs and coding RNAs. Expressed lncRNAs have an activating histone modification profile similar to that of protein-coding genes, with slightly excess levels of both silencing (H3K27me3) and activating (H3K36me3) marks in lncRNAs [24]. Although many lncRNAs are polyadenylated, they are significantly enriched in nonpolyadenylated transcripts compared with coding RNAs [24]. As expected, the lncRNAs have significantly lower protein potentials compared with mRNAs. Interestingly, they are biased toward two-exon transcripts and predominately localized in the chromatin and nucleus, while coding RNAs are predominantly observed in the cytosol; only 6% of them have a 2-exon structure.

In order to identify the function and targets of lncRNAs, Derrien *et al.* [24] investigated the correlation between lncRNAs and coding RNAs; both positive and negative correlations were found. Overall, the number of positive correlations was larger than the negative correlations. Compared with trans-acting lncRNAs, cis-acting lncRNAs showed more positive correlations. Interestingly, lncRNAs that intersected protein-coding exons in the antisense orientation showed a strong pattern of coexpression [24]. They also checked sequence conservation by BLASTing human lncRNAs against all available mammalian genomes. Around 30% (n = 4,546) of lncRNAs appeared to have arisen in the primate lineage, and 0.7% (n = 101) of them appeared to be human-specific [24].

RNA splicing

Recent studies suggest that many pre-mRNA processing events are cotranscriptional [25-27]. RNA imaging found that splicing can follow the completion of transcription [28]. Tilgner *et al*. [29] investigated cotranscriptional splicing by interrogating RNA fractions from several cellular compartments in K562 cells. They found that only a tiny fraction of exons were found to be surrounded completely by an unspliced intron in chromatin-associated RNA. This suggests that splicing is already occurring during transcription [18, 29]. The strong enrichment of spliceosomal small nuclear RNAs (snRNAs) in the chromatin-associated fraction compared with other fractions also supports cotranscriptional splicing. These observations confirm the idea that chromatin structure could play a role in splicing [7, 30-34]. However, for alternative exons and lncRNAs, splicing tends to occur later and might remain unspliced in some cases [29].

RNA editing

Li *et al*. [35] reported 10,210 exonic sites in the human

genome where an RNA sequence did not match with the DNA sequence, suggestive of RNA editing. However, there have been debates about whether their observations occurred by sequencing error, gene duplication, mapping error, or read-end misalignment [36-38].

Park *et al*. [39] developed a pipeline to filter sequencing artifacts in identifying RNA editing. They found that the majority of non-A-to-G variants came from incorrect read mapping across splice junctions. Most of the edits they found were A-to-G(I) variants, which corresponds with recent observations [40] but differs from Li *et al*.'s report [35] of a substantial number of noncanonical single nucleotide variant edits in the RNA of human lymphoblastoid cells. Most A-to-G(I) edits were located in introns and untranslated regions, with only a fraction of sites reproducibly edited across multiple cell lines.

Pseudogenes

Pseudogenes have been considered nonfunctional sequences of genomic DNA that lost their coding potential due to disruptive mutations, such as frameshifts and premature stop codons [41-44]. Recent studies have shown that pseudogenes can regulate their parent genes [45-49]. Using manual annotation, with the assistance of computational pipelines, GENCODE created a database called Pseudogene Decoration Resource (psiDR). psiDR provides a variety of information on 11,216 peudogenes, including transcription activity, chromatin features, functional genomics, and evolutionary constraints [50]. Transcribed pseudogenes show enhanced chromatin accessibility and enrichment with histone marks, although they are lower than those of coding genes. The majority of pseudogenes contains no or very few transcription factor binding sites (TFBSs), but the differences between the number of TFBSs associated with transcribed and nontranscribed pseudogenes are significant.

Small RNAs

GENCODE also categorized 7,054 small RNAs into 2,756 micro RNAs (miRNAs), snRNAs, small nucleolar RNAs (snoRNAs), and transfer RNAs (tRNAs). They found that miRNAs and tRNAs were abundant in cytosol, snoRNAs were in the nucleus, and snRNAs were in both the nucleus and cytosol. snRNAs were found abundantly in the chromatin-associated RNA fractions, which further supports predominant splicing during transcription.

Enhancer RNAs

RNAs at enhancers (eRNAs) were first characterized by observing transcription activities at the promoter-distal CREB binding protein-binding sites in mouse cortical neurons [51]. The bidirectional property of eRNAs and their association with gene expression were further studied using nascent RNAs from global run-on sequencing (GROseq) data [52, 53]. ENCODE used RNA assays to detect transcription activity. Besides the bidirectional property, ENCODE identified transcriptional initiation using cap analysis gene expression (CAGE) [54] signals. Interestingly, they found polyadenylated eRNAs, although most eRNAs were prevalent in the nonpolyadenylated form. They also observed that histone marks associated with eRNAs were the factors for transcriptional initiation and elongation: H3K27ac, H3K79me2, and RNA polymerase. These lines of evidence suggest regulatory functions of eRNAs.

Table 3. Summary of data collected for modeling gene expression using chromatin features

Description	Classification	Detail
Chromatin	Methylation	H3K4me1, H3K4me2, H3K4me3, H3K27me3, H3K36me3, H3K79me2, H3K9me1, H3K9me3, H4K20me1
	Acetylation	H3K9ac, H3K27ac
	Others	H2A.Z, DNase I hypersensitivity
Cell lines	Tier 1	K562, GM12878, H1-hESC
	Tier 2	HepG2, HeLa-S3, NHEK, HUVEC
Cellular compartments	Nucleus	Further isolated into nucleolus, nucleoplasm, and chromatin in selected cell lines
	Cytosol	
	Whole cell	
RNA extractions	PolyA+	
	PolyA−	
RNA sequencing technology	CAGE	Cap analysis of gene expression [54]
	RNA-PET	RNA paired end tag [55]
	RNA-seq	Whole transcript coverage [56]

Predicting Gene Expression by Chromatin Features

Enabled by the unprecedented volume of data generated by the ENCODE project, Dong *et al.* [57] performed a very interesting study that attempted to predict gene expression from chromatin features using machine learning techniques. As usually practiced, a machine learning study is composed of a series of procedures that typically involve data collection, data representation, model building, and testing. Using machine learning terminology, the response variables here are gene expression patterns that are predicted/modeled, while the predictors or features are various chromatin measures. Chromatin data were collected from the ENCODE project with 11 chromatin modifications, one histone variant, and DNase I hypersensitivity, all mapped in seven cell lines (Table 3). Gene expression data were from different cellular compartments, using two different RNA isolation approaches and sequenced with different technologies (Table 3). With such diversity of RNA sources, Dong *et al.* [57] answered not only the general question of whether gene expression can be predicted with satisfactory accuracy but also the questions of whether different RNA sources that are sequenced by different technologies can be predicted differently using the same chromatin features.

The gene expression data are relatively easy to represent. They were separated into two classes: transcriptional start site (TSS)-based and transcript-based (Tx-based). TSS-based expression data are read counts within a 101-bp window centered on the TSS, which measures transcriptional initiation. Tx-based expression data are summarized read counts from the whole transcript, which measures transcriptional elongation. However, the representation of chromatin data seems to be tricky and requires further research. Dong *et al.* [57] used a strategy called "bestbin," which considers the chromatin signals across the entire gene body, including 2-kb flanking regions. It basically segregates each genic region into equal bins of 100 bp and summarizes the chromatin signals within each. A training dataset was used to identify the bin that correlates most with gene expression, and the learned parameter values were applied to testing data. Other strategies [58, 59] are possible, but the "bestbin" strategy was found to be superior [57].

RNA sequencing data are known to contain very little or no background noise, with a large portion of the genes having 0 read counts. Therefore, the response variables become a mixture of a 0 component and a positive counting component. Neither a classifier nor regression method seems to be able to capture the variability of both. To deal with the challenge, a two-step approach was used by [57], so that a classifier first categorized genes as "expressed" or "unexpressed," Then, a regression method was used to predict the expression levels of the expressed genes. The final prediction is the product of the classifier and the regression method.

To test the performance of their approach, each dataset was separated into a training set and a testing set. On the training set, the "best bin" and a few other parameters were determined. After that, a 10-fold crossvalidation was performed on the testing set to evaluate the model. AUC [60] was used to represent the accuracy of the classifier. Two criteria were used to represent the accuracy of the regression method. Pearson correlation coefficient (PCC) was used to measure the similarity between the predicted value and experimental value. Root mean square error was used to measure the disparity between the predicted value and experimental value.

Overall, the two-step model achieved very satisfactory performance, with a PCC > 0.9 for a number of datasets and a PCC > 0.8 for 71% of the whole data. Looking at the two steps separately, the AUC can be as high as 0.95 for the classifier, and the PCC can be as high as 0.77 for the regression method when predicting CAGE-measured polyA+ cytosolic RNA expression in K562 cells. Similar performance was achieved in other datasets. It was also found that H3K9ac and H3K4me3 are the most important predictors for the classifier, strengthening their roles as activation marks at TSSs. In contrast, H3K79me2 and H3K36me3 are the most important predictors for the regression methods, strengthening their roles as elongation marks at gene bodies. These findings show that the two-step model not only improved the accuracy of prediction but also enabled the identification of the chromatin features that are associated with different transcriptional roles.

Discussion

During the past decade, the ENCODE project has evolved into a genomewide scale, and the dataset it generated has expanded in quantity as well as in scope. The ENCODE project has provided a global view about the human transcriptome and most noticeably found that the transcribed region of the human genome is more abundant than we previously thought. This finding significantly reduced the so-called intergenic regions, as defined in the traditional sense. The quantitative measurement of RNA species in several cellular compartments as well as their polyadenylation provided a comprehensive view of RNA generation. In this review, we have revisited the characteristics of both coding and noncoding transcripts in association with their structures and locations in cells.

Besides the human transcriptome and the associated chromatin modification data that we discussed, the ENCODE consortium also mapped transcription factor binding sites and their associated DNA motifs, as well as DNA methylation and long-range chromosomal interactions [4]. In parallel, the Roadmap Epigenomics Project (http://www.roadmapepigenomics.org) and International Human Epigenome Consortium (http://www.ihec-epigenomes.org) have been accumulating data of a similar scale to understand the human genome in other tissues and conditions. It is remarkable that the ENCODE data altogether have associated more than 80% of the human genome with some type of biochemical function so far, and the coverage will continue to increase as we map additional protein-DNA interactions in the near future. It has now become very clear that so-called "junk DNA" is not evolutionarily vestigial but has specific structural or biochemical functions.

While data generation has been a major goal of ENCODE, the need to integrate the current datasets is becoming more and more important. Computational approaches have been developed to exploit the ENCODE data at to a fuller potential. For instance, chromatin features were used to model gene expression [57, 61]; integrative methods were developed to annotate genomes [62-64]; visualization tools were developed to investigate epigenomic regulation at a global scale (also see ngs.plot at https://code.google.com/p/ngsplot/) [65, 66]; and large regulatory networks were reconstructed, based on TFs and DNaseI footprinting [67, 68]. The network-based approach as well as the chromosomal interactions [69] provided novel angles in studying gene regulation at higher levels. New approaches to integrate the large amount of data to provide new biological insights are on the horizon.

References

1. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004;306:636-640.
2. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, *et al*. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007;447:799-816.
3. ENCODE Project Consortium, Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, *et al*. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* 2011;9:e1001046.
4. ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, *et al*. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57-74.
5. Strahl BD, Allis CD. The language of covalent histone modifications. *Nature* 2000;403:41-45.
6. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, *et al*. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 2009;459:108-112.
7. Andersson R, Enroth S, Rada-Iglesias A, Wadelius C, Komorowski J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res* 2009;19:1732-1741.
8. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of alternative splicing by histone modifications. *Science* 2010;327:996-1000.
9. Kouzarides T. Chromatin modifications and their function. *Cell* 2007;128:693-705.
10. Zhang ZD, Paccanaro A, Fu Y, Weissman S, Weng Z, Chang J, *et al*. Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. *Genome Res* 2007;17:787-797.
11. Johnson DS, Li W, Gordon DB, Bhattacharjee A, Curry B, Ghosh J, *et al*. Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res* 2008;18:393-403.
12. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! *Genome Biol* 2011;12:125.
13. Ho JW, Bishop E, Karchenko PV, Nègre N, White KP, Park PJ. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics* 2011;12:134.
14. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;10:669-680.
15. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, *et al*. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 2012;22:1813-1831.
16. Chen Y, Negre N, Li Q, Mieczkowska JO, Slattery M, Liu T, *et al*. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat Methods* 2012;9:609-614.
17. Egelhofer TA, Minoda A, Klugman S, Lee K, Kolasinska-Zwierz P, Alekseyenko AA, *et al*. An assessment of histone-modification antibody quality. *Nat Struct Mol Biol* 2011;18:91-93.
18. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, *et al*. Landscape of transcription in human cells. *Nature* 2012;489:101-108.
19. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, *et al*. Ensembl 2012. *Nucleic Acids Res* 2012;40:D84-D90.
20. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, *et al*. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 2009;458:223-227.
21. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, *et al*. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* 2010;28:503-510.
22. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. *Nat Rev Genet* 2009;10:155-159.
23. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, *et al*. Antisense transcription in the mammalian transcriptome. *Science* 2005;309:1564-1566.

24. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, *et al*. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 2012;22:1775-1789.
25. Reed R. Coupling transcription, splicing and mRNA export. *Curr Opin Cell Biol* 2003;15:326-331.
26. Kornblihtt AR, de la Mata M, Fededa JP, Munoz MJ, Nogues G. Multiple links between transcription and splicing. *RNA* 2004; 10:1489-1498.
27. Listerman I, Sapra AK, Neugebauer KM. Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. *Nat Struct Mol Biol* 2006;13:815-822.
28. Vargas DY, Shah K, Batish M, Levandoski M, Sinha S, Marras SA, *et al*. Single-molecule imaging of transcriptionally coupled and uncoupled splicing. *Cell* 2011;147:1054-1065.
29. Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, *et al*. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* 2012;22:1616-1625.
30. Hon G, Wang W, Ren B. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput Biol* 2009;5:e1000566.
31. Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet* 2009;41:376-381.
32. Nahkuri S, Taft RJ, Mattick JS. Nucleosomes are preferentially positioned at exons in somatic and sperm cells. *Cell Cycle* 2009;8:3420-3424.
33. Spies N, Nielsen CB, Padgett RA, Burge CB. Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell* 2009;36:245-254.
34. Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcárcel J, *et al*. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* 2009;16:996-1001.
35. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, *et al*. Widespread RNA and DNA sequence differences in the human transcriptome. *Science* 2011;333:53-58.
36. Pickrell JK, Gilad Y, Pritchard JK. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* 2012;335:1302.
37. Kleinman CL, Majewski J. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* 2012;335:1302.
38. Lin W, Piskol R, Tan MH, Li JB. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* 2012;335:1302.
39. Park E, Williams B, Wold BJ, Mortazavi A. RNA editing in the human ENCODE RNA-seq data. *Genome Res* 2012;22: 1626-1633.
40. Peng Z, Cheng Y, Tan BC, Kang L, Tian Z, Zhu Y, *et al*. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol* 2012; 30:253-260.
41. Mighell AJ, Smith NR, Robinson PA, Markham AF. Vertebrate pseudogenes. *FEBS Lett* 2000;468:109-114.
42. Harrison PM, Echols N, Gerstein MB. Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. *Nucleic Acids Res* 2001;29:818-830.
43. Echols N, Harrison P, Balasubramanian S, Luscombe NM, Bertone P, Zhang Z, *et al*. Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. *Nucleic Acids Res* 2002;30: 2515-2523.
44. Balakirev ES, Ayala FJ. Pseudogenes: are they "junk" or functional DNA? *Annu Rev Genet* 2003;37:123-151.
45. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 2010;465: 1033-1038.
46. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, *et al*. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* 2008;453:534-538.
47. Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, *et al*. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* 2008;453:539-543.
48. Sasidharan R, Gerstein M. Genomics: protein fossils live on as RNA. *Nature* 2008;453:729-731.
49. Guo X, Zhang Z, Gerstein MB, Zheng D. Small RNAs originated from pseudogenes: cis- or trans-acting? *PLoS Comput Biol* 2009;5:e1000449.
50. Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, *et al*. The GENCODE pseudogene resource. *Genome Biol* 2012;13: R51.
51. Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, *et al*. Widespread transcription at neuronal activity-regulated enhancers. *Nature* 2010;465:182-187.
52. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 2008;322:1845-1848.
53. Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, *et al*. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* 2011;474:390-394.
54. Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, *et al*. CAGE: cap analysis of gene expression. *Nat Methods* 2006;3:211-222.
55. Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, *et al*. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods* 2005;2: 105-111.
56. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;10:57-63.
57. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, *et al*. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol* 2012;13:R53.
58. Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A* 2010;107:2926-2931.
59. Costa IG, Roider HG, do Rego TG, de Carvalho Fde A. Predicting gene expression in T cell differentiation from histone modifications and transcription factor binding affinities

by linear mixture models. *BMC Bioinformatics* 2011;12 Suppl 1:S29.
60. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2006;27:861-874.
61. Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res* 2012;22:1711-1722.
62. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, *et al*. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* 2013;41:827-841.
63. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, *et al*. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 2011;473:43-49.
64. Won, KJ, Zhang X, Wang T, Ding B, Raha D, Snyder M, *et al*. Comparative annotation of functional regions in the human genome using epigenome data. *Nucleic Acids Rese* 2013;41: 4423-4432.
65. Shin H, Liu T, Manrai AK, Liu XS. CEAS: cis-regulatory element annotation system. *Bioinformatics* 2009;25:2605-2606.
66. Ye T, Krebs AR, Choukrallah MA, Keime C, Plewniak F, Davidson I, *et al*. seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res* 2011;39:e35.
67. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, *et al*. Architecture of the human regulatory network derived from ENCODE data. *Nature* 2012;489:91-100.
68. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, *et al*. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 2012;489: 83-90.
69. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature* 2012;489: 109-113.