CellPress

Review

Enhanced Identification of Transcriptional Enhancers Provides Mechanistic Insights into Diseases

Yasuhiro Murakawa,[1,2,6,*] Masahito Yoshihara,[2,3,6] Hideya Kawaji,[1,2,4] Miki Nishikawa,[1] Hatem Zayed,[5] Harukazu Suzuki,[2] FANTOM Consortium,[1,2,3,4] and Yoshihide Hayashizaki[1]

Enhancers are distal *cis*-regulatory DNA elements that increase the expression of target genes. Various experimental and computational approaches including chromatin signature profiling have been developed to predict enhancers on a genome-wide scale, although each method has its advantages and disadvantages. Here we overview an emerging method to identify transcribed enhancers at exceedingly high nucleotide resolution based on enhancer RNA transcripts captured by Cap Analysis of Gene Expression (CAGE) technology. We further argue that disease-causative regulatory mutations at enhancers are increasingly recognized, emphasizing the importance of enhancer identification in functional and clinical genomics including, but not limited to, genome-wide association studies (GWASs) and cancer genomics studies.

Regulation of Gene Expression by Enhancers

Spatiotemporal control of gene expression is of critical importance for cellular differentiation, organogenesis, and homeostasis in multicellular organisms and dysregulation of gene expression is linked to many diseases (reviewed in [1,2]). Regulation of gene expression is a multilayered process [3] and an initial step is the synthesis of RNAs. Besides core promoter sequences in the immediate proximity of transcription start sites (TSSs), which recruit general transcription factors (TFs) and initiate RNA polymerase II (RNAPII)-mediated transcription, other *cis*-acting elements such as proximal promoters, enhancers, and silencers cooperatively modulate basal transcription from the core promoters (reviewed in [4]). Among these, enhancers are small segments of promoter-distal *cis*-regulatory DNA regions that significantly enhance the expression of target genes independent of location or orientation with respect to the target genes (reviewed in [5,6]).

An enhancer was first described in the genome of simian virus 40 (SV40) in 1981 by Banerji *et al.* [7] and Moreau *et al.* [8], where a 72-bp repeated sequence located upstream of the SV40 early region significantly increased the ectopic expression of a reporter gene. Remarkably, a non-viral enhancer was discovered in 1983 within a mouse immunoglobulin heavy chain gene [9–11], followed by studies documenting many enhancers in various organisms. Thus, an enhancer was originally defined by its functionality in enhancing the transcription of target genes.

Trends

Various experimental and computational approaches have been developed to predict enhancers on a genome-wide scale, although each method has its advantages and disadvantages.

Cap analysis of gene expression (CAGE) identifies transcribed enhancers at exceedingly high nucleotide resolution by detecting enhancer RNAs (eRNAs).

Disease-associated SNPs and recurrent somatic cancer mutations are identified within enhancers. These variants might alter enhancer activities and contribute to pathogenesis, highlighting the importance of enhancer identification in various clinical settings.

[1]RIKEN Preventive Medicine and Diagnosis Innovation Program, Wako, Saitama 351-0198, Japan
[2]RIKEN Center for Life Science Technologies (CLST), Division of Genomic Technologies (DGT), Kanagawa 230-0045, Japan
[3]Department of Ophthalmology, Osaka University Graduate School of Medicine, Osaka 565-0871, Japan
[4]RIKEN Advanced Center for Computing and Communication, Preventive Medicine and Applied Genomics Unit, Kanagawa 230-0045, Japan

[5]Department of Health Sciences, Biomedical Program, Qatar University, PO Box 2713, Doha, Qatar
[6]These authors contributed equally to this work

*Correspondence: yasuhiro.murakawa@riken.jp (Y. Murakawa).

Enhancers contain binding sites for TFs that interact with coactivators including the histone acetyltransferases p300/CREB-binding protein (CBP) (reviewed in [12,13]). An enhancer is then brought into close proximity with its target promoter through chromatin looping, facilitated by Mediator, which associates with cohesin and the cohesin loading factor [14]. Regions bound by TFs are typically depleted of nucleosomes and are sensitive to cleavage by deoxyribonuclease I (DNase I) [15]. These nucleosome-depleted regions (NDRs) are flanked by regions marked with specific histone modifications such as histone H3 lysine 4 monomethylation (H3K4me1) and H3K27 acetylation (H3K27ac). H3K4me1 is associated with inactive, poised, and active enhancers whereas H3K27ac is more specifically associated with active enhancers (reviewed in [16,17]).

Super-enhancers have been proposed to describe groups of putative enhancers clustering in close genomic proximity. These can span exceptionally large genomic regions with strong enrichment for the binding of TFs and Mediator [18–20]. Intriguingly, super-enhancers are often located near genes related to cell type-specific function and are enriched for sequence motifs of cell type-specific master TFs. However, the term super-enhancer has been used in many studies without a clear definition (reviewed in [21]).

Various Approaches to Identify Enhancers

With an increasingly clearer picture of how enhancers work, various computational and experimental approaches have been developed to identify enhancers (Table 1).

Comparative genomic analyses on conserved noncoding sequences or TF-binding motifs successfully identified a fraction of novel enhancers [22–24] (reviewed in [25,26]). Candidate enhancer elements together with their *in vivo* validation experiments are available through VISTA Enhancer Browser[i] [27]. With the advent of next-generation sequencing technology, many high-throughput experimental approaches were developed to predict enhancers on a genome-wide scale. ChIP-seq of TFs predicts a subset of putative enhancers (reviewed in [13]) whereas ChIP-seq of p300 covers enhancers more ubiquitously [28]. High-throughput profiling of DNase I hypersensitive sites (DHSs) allows identification of enhancers [15], although DHSs also include other regulatory DNA regions such as promoters, insulators, and silencers. More recently, an assay for transposase-accessible chromatin using sequencing (ATAC-seq) was developed as a rapid and sensitive alternative method for examining chromatin accessibility [29] (reviewed in [30]). In addition, ChIP-seq of histone modifications enables genome-wide prediction of putative enhancers and combinatorial analysis of distinct histone marks allows determination of enhancer activation (reviewed in [16,17]).

The Encyclopedia of DNA Elements (ENCODE) Project has contributed enormously to building a comprehensive map of functional elements in the human genome[ii] [31] (Box 1). They have provided 457 ChIP-seq datasets on more than 119 TFs in many human cell lines and these data are available on Factorbook[iii] [32]. Furthermore, Ernst *et al.* annotated 15 chromatin states by ChIP-seq of nine distinct histone modifications across nine cell types. They defined regions with high H3K4me1 and H3K27ac levels as strong enhancers and regions with high H3K4me1 but low H3K27ac levels as weak or poised enhancers [33].

The term enhancer is ambiguously defined and different methods identify 'enhancers' by capturing different features or aspects (e.g., NDRs, TF-binding sites, histone marks surrounding NDRs). Moreover, each method has its advantages and disadvantages. NDRs, which can be identified by DNase-seq or ATAC-seq, are observed at diverse types of regulatory elements in addition to enhancers [30]. Similarly, TFs bind to a broad spectrum of regulatory elements [13]. In addition, ChIP-seq data suffer from the intrinsic issue of limited base resolution. This issue arises from sonication-based chromatin fragmentation followed by size selection of 200–400-bp

Table 1. Summary of Methods for Genome-Wide Identification of Enhancers

Method	Advantages	Limitations	Refs
Computational Analysis of Conservation and TF-Binding Motifs			Reviewed in [26]
Conservation analysis of noncoding elements between different species	Novel conserved DNA segments or TF-binding motifs could be discovered.	Many other regulatory elements could be included. Not all enhancers are conserved.	[22–24]
Combinatorial analysis of TF-binding motif scan and conservation	Easily applicable to many TFs with known motifs.	High-false positive rate.	[84]
Regulator Binding			Reviewed in [13]
ChIP-seq of TFs	Occupancy of key TFs predicts enhancers.	Key TFs of the target cells must be known. ChIP-grade antibodies are not available for all TFs. Difficult to distinguish enhancers and promoters. Do not cover all enhancers.	[85,86]
ChIP-seq of transcriptional coactivators such as p300	p300-binding sites are reported to predict enhancers generally.	Difficult to distinguish active and poised enhancers.	[28,87]
Chromatin Accessibility			Reviewed in [30]
DNase-seq	Extensively used for various types of cells.	Includes diverse types of regulatory elements such as promoters, insulators, and silencers besides enhancers.	[88]
Formaldehyde-assisted isolation of regulatory elements (FAIRE)-seq			[89]
ATAC-seq	Applicable to small numbers of cells. Short library preparation time.		[29]
Histone Modifications			Reviewed in [17]
ChIP-seq profiling of histone modifications (e.g., H3K4me1, H3K27ac)	ChIP-grade antibodies are broadly available and applicable for many species.	Broad peaks make it difficult to predict enhancers with high nucleotide resolution.	[33,90]
Chromosomal Interaction			Reviewed in [91]
Chromatin interaction analysis with paired-end tag sequencing (ChIA-PET)	ChIA-PET of RNAPII predicts enhancers as well as enhancer–promoter interactions.	Spatial proximity does not necessarily reflect functional regulatory relationships.	[92]
Functional Enhancer Screening by Reporter Assay			Reviewed in [6]
Massively parallel reporter assays (MPRAs)	Directly assess enhancer activities in parallel by sequencing the transcribed reporter containing heterologous barcodes.	Not optimal for genome-wide screening.	[93]
Self-transcribing active regulatory region (STARR)-seq	Directly assesses genome-wide enhancer activities with a high detection rate by sequencing the transcribed enhancers themselves.		[94]
eRNA-Based Detection			Reviewed in [38]
RNA-seq	Expression levels of eRNAs and their nearby genes are quantified simultaneously.	Low nucleotide resolution. Difficult to detect eRNAs, which are typically lowly expressed. Detects only transcribed enhancers (not all enhancers are transcribed).	[36]

Table 1. (continued)

Method	Advantages	Limitations	Refs
Chromatin-associated RNA (ChromRNA)-seq	Detects unstable transcripts including eRNA by enriching chromatin fraction.	Detects only transcribed enhancers, which are more likely to be active.	[95]
GRO-seq PRO-seq	Detects unstable transcripts including eRNAs.	Requires elaborate *in vitro* experimental procedures. Detects only transcribed enhancers.	[53,55]
NET-seq	Detects unstable transcripts including eRNAs. Detects 3′ ends of RNAPII-bound eRNAs at nucleotide resolution.	Detection of 3′ ends of eRNAs makes it difficult to pinpoint enhancer region. Detects only transcribed enhancers.	[58]
CAGE	Detects the TSSs of eRNAs at nucleotide resolution. Expression levels of eRNAs and their nearby genes are quantified simultaneously.	Requires large sample size to detect lowly expressed eRNAs. Detects only transcribed enhancers.	[41]

fragments during library preparation [34]. Histone modification-based prediction of enhancers has been widely used, but even when the analyses are done using multiple histone marks it sometimes still suffers from false-positive/negative identifications [35]. Furthermore, 'enhancers' identified using histone modification ChIP-seq analysis typically contain much wider regions because H3K27ac, for example, occurs broadly surrounding NDRs. This complicates enhancer identification at high resolution.

Strikingly, seminal work by Kim *et al.* demonstrated widespread bidirectional RNA transcription originating from thousands of enhancers [36]. In the following sections, we overview recent findings and further discuss emerging methods for enhancer identification based on RNA transcription at enhancers.

Widespread Transcription at Enhancers

Transcription at active enhancers was first reported in the locus control region (LCR) of the *beta-globin* gene [37] and recent technological advances in high-throughput sequencing revealed that transcription at enhancers is a widespread phenomenon (reviewed in [38]). In 2010, Kim *et al.* sequenced rRNA-depleted total RNAs from mouse cortical neurons and discovered bidirectional RNA transcription originating from thousands of chromatin state-defined enhancers. They referred to these RNAs as enhancer RNAs (eRNAs) [36]. Importantly, this finding highlighted that enhancers are not just sites bound by TFs but are also the sites where active transcription occurs. Consistently, enhancers are also bound by transcription pre-initiation complexes, including RNAPII [39].

Box 1. ENCODE Project

The ENCODE Project was launched by the National Human Genome Research Institute (NHGRI) in 2003 with the aim of annotating functional elements in the human genome. They have developed many types of methods and performed numerous high-throughput experiments combined with computational analyses to identify functional elements across the human genome. The initial analysis of 1640 datasets obtained from 24 types of experiment in 147 cell types was reported in 2012, revealing that 80.4% of the human genome might have functionality in at least one cell type [31]. Various genome-wide studies across a wide range of cell types were performed, including RNA profiling (e.g., CAGE, RNA-seq), chromatin structure (e.g., DNase-seq, histone ChIP-seq), TF-binding sites (ChIP-seq), and DNA methylation sites (bisulfite sequencing). These data are still being updated and the University of California, Santa Cruz (UCSC) has made these datasets available on the public Genome Browser[iv,v] [96].

In general, eRNAs are capped at their 5′ end similarly to mRNAs but are not polyadenylated or spliced. eRNAs tend to be more bidirectionally transcribed, albeit with a varying degree of bidirectionality [36,40,41]. Importantly, eRNAs are susceptible to exosome-mediated degradation and are expressed at very low levels [40,42] (features of eRNAs are reviewed in [38]).

The functionality of eRNA was reported by Melo *et al.* [43]. They showed in a human breast cancer cell line that p53-bound regions located distantly from known p53 target genes generated eRNAs in a p53-dependent manner. Notably, siRNA-mediated knockdown of these eRNAs reduced the p53-dependent induction of the target mRNAs. Similarly, Li *et al.* reported that, in human breast cancer cells, stimulation of estrogen receptor alpha (ERα) with 17β-estradiol (E2) caused a global increase of eRNA expression at enhancers located adjacent to E2-upregulated coding genes. Specific inhibition of eRNA using siRNA or locked nucleic acid antisense oligonucleotides (LNAs) abrogated the ligand-dependent induction of target genes [44]. Roles of eRNA in the induction of target genes have been reported in many biological contexts [45–47]. However, a detailed molecular mechanism of eRNA function remains elusive.

Functional Annotation of the Mammalian Genome 5 (FANTOM5) Revealed a High-Resolution View of Transcribed Enhancers

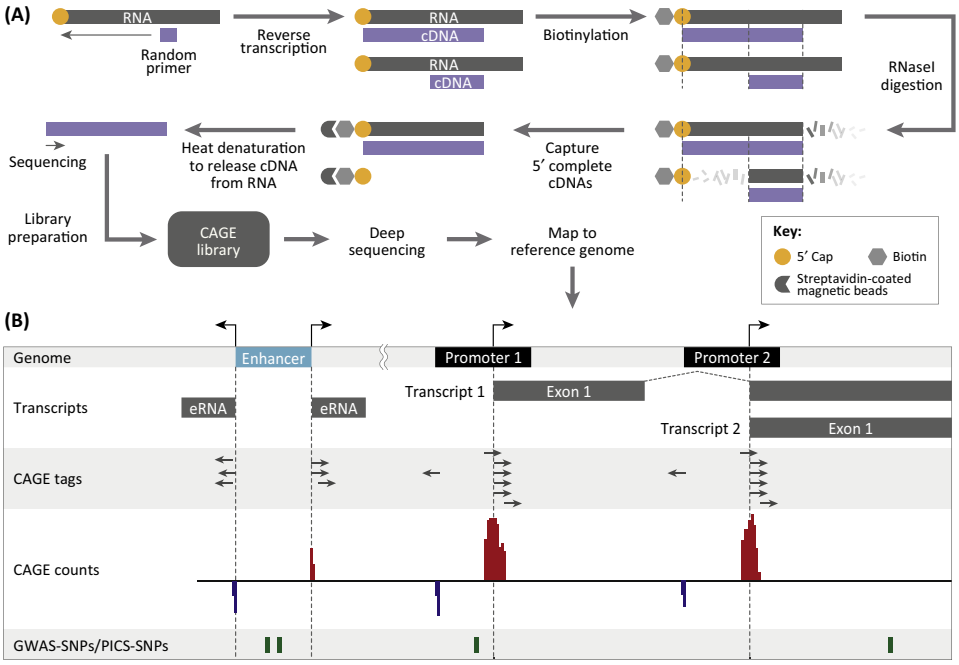
Inspired by recent findings that transcription occurs at active enhancers, transcribed enhancers were identified on a genome-wide scale by CAGE at exceedingly high nucleotide resolution in the FANTOM5 project (Box 2) [41].

CAGE is a technique that was originally developed by RIKEN for the genome-wide detection and quantification of TSSs (Figure 1) [48]. Short sequences (typically 25–27 nt bp) from the capped 5′ ends of transcripts are selectively subjected to high-throughput sequencing (a HeliScope single-molecule sequencer was used in the FANTOM5 project), followed by mapping to the reference genome. Although CAGE does not provide information on the complete architectures of transcripts, it uses all of the sequencing capacity to deep sequence the 5′ ends of transcripts, allowing robust quantification of the 5′ ends of transcripts that exist at low levels (Figure 1). CAGE has been an essential tool used in international projects including ENCODE [31] (Box 1) and FANTOM (Box 2).

Remarkably, 43 011 putative transcribed enhancers were identified based on the presence of bidirectional transcripts in high-depth FANTOM5 CAGE data collected from 808 human samples encompassing 432 primary cells, 135 tissues, and 241 cell lines [41]. Enhancer-associated forward and reverse strand transcription initiation sites were separated by around 180 bp on average and overlapped with nucleosome boundaries. Enhancers determined by CAGE

Box 2. FANTOM Project

FANTOM is an international research consortium initially established in 2000 to achieve annotation of mouse full-length cDNAs[vi]. A recent review article [97] overviews the FANTOM projects. Briefly, in FANTOM1 many original experimental technologies as well as computational tools were developed to annotate ~20 000 mouse full-length cDNAs. These annotations also contributed to gene prediction of the human genome. In FANTOM2, the full-length cDNA collection was expanded to 60 770. The collection unexpectedly contained many cDNAs that did not appear to encode proteins. In FANTOM3, a new technology, CAGE, was developed to enable high-throughput identification of TSSs as well as quantification of their adjacent promoter activities. CAGE was applied to produce a comprehensive promoter map for human and mouse, revealing different classes of promoter architecture as well as numerous noncoding RNAs. In FANTOM4, the cellular transcriptional network required for cellular differentiation was studied using CAGE technology. Recently, the FANTOM5 phase 1 project performed CAGE using HeliScope single-molecule sequencing technology across 975 human and 399 mouse samples including primary cells, tissues, and cancer cell lines to comprehensively identify cell type-specific promoters and enhancers. The FANTOM5 phase 2 project studied the dynamics of enhancers and promoters in 19 human and 14 mouse time-course experiments during stimulus-dependent activation and cellular differentiation into specialized cell types and surprisingly discovered that enhancer transcription preceded the activation of the target promoters [98].



Trends in Genetics

Figure 1. Overview of Cap Analysis of Gene Expression (CAGE) Library Preparation and Data Analysis. (A) Schematic overview of CAGE library preparation. First, cDNAs are reverse transcribed using random primers. Then, 5′ caps of RNAs are biotinylated. After RNase I digestion of single-stranded RNAs, biotinylated RNA/cDNA hybrid molecules are captured by streptavidin-coated magnetic beads. Subsequently, the cDNAs are released from RNAs to prepare the CAGE library. Finally, short sequences from the capped 5′ ends (CAGE tags) are obtained from the CAGE library by deep sequencing. (B) Overview of CAGE data analysis for the quantification of promoter and enhancer activities. By mapping CAGE tags to the reference genome, transcription start sites are detected at nucleotide resolution. Promoter activities can be inferred based on the CAGE tag counts [99]. Simultaneously, CAGE allows the determination of enhancers and quantification of their eRNA expression levels based on the bidirectional capped eRNAs [41]. Red and blue signals indicate CAGE signals in the forward and reverse direction, respectively.

represent regions between forward and reverse strand eRNA initiation sites, thereby providing exceedingly high nucleotide resolution. However, eRNAs could be reliably detected only by using pools of libraries with high sequencing coverage, due to their much lower level of expression. Additionally, bidirectionality of eRNA expression is typically observed only in a meta-analysis, which could introduce false-positives and false-negatives when predicting enhancers. Furthermore, it is sometimes difficult to distinguish between enhancers and promoters because they have many similarities, as discussed in a recent review [49]. However, most CAGE-defined enhancers overlapped with H3K4me1/H3K27ac and DHSs and CAGE-defined transcribed enhancers showed much higher success rates in reporter validation assays than untranscribed enhancers.

The FANTOM5 CAGE library collection allowed systematic analysis of enhancer usage across a large variety of cell types and tissues in the human body. Notably, enhancers tend to be more specifically transcribed in a much smaller subset of samples compared with mRNA transcripts, which is in line with the expected roles of enhancers in lineage specification.

It has been a major challenge to link enhancers to the genes they regulate because enhancers can be located at any distance from the target genes. A hint came from previous reports showing that eRNA expression correlates with the expression of target genes [36,43,44]. Uniquely, the FANTOM5 CAGE data atlas enables direct pairwise correlations between transcription of the

enhancer and that of putative target genes across a broad panel of cells, revealing putative target genes for CAGE-defined enhancers. On average, a TSS of a protein-coding gene was predicted to be targeted by 4.9 enhancers[vii].

Furthermore, the FANTOM5 CAGE expression atlas, which contains data from hundreds of cell types, is a valuable resource for future investigation. Yao *et al.*, taking advantage of the FANTOM dataset, identified a robust set of eRNAs and constructed networks with eRNA–gene coexpression interactions across human fetal brain and multiple adult brain regions [50].

Overview of Methods to Detect eRNAs

There are several methods that could be applied to detect eRNAs. Originally, eRNAs were detected by total RNA sequencing [36]. By contrast, CAGE defines the 5′ end of transcripts, which makes it possible to pinpoint the exact position of eRNA initiation, thus providing an unprecedented enhanced view of transcribed enhancers [41]. Besides CAGE, several other methods detect the 5′ end of transcripts, including TSS-seq [51] and paired-end analysis of TSSs (PEAT) [52]. However, eRNAs are expressed at very low levels and could be readily detected only by high sequencing coverage. Therefore, detection methods need to be quite sensitive to accurately predict enhancers.

To address this issue, enrichment of nascent transcripts could be performed. Unstable eRNAs sensitive to exosome-mediated degradation were successfully detected by global run-on sequencing (GRO-seq) [53], in which nascent transcripts containing labeled nucleotides are purified and sequenced [54]. However, GRO-seq has limited base resolution, and although precision nuclear run on and sequencing (PRO-seq) achieved higher base resolution [55] both methods require elaborate experimental procedures such as pre-isolation of nuclei and *in vitro* nuclear run-on assays. These procedures could introduce experimental variability and bias. In addition, start-seq detects the 5′ ends of short capped RNAs derived from stalled polymerase [56]. Scruggs *et al.* applied this method to murine macrophages and discovered that a subset of antisense TSSs displayed enhancer-like chromatin signatures such as high levels of H3K4me1, H3K27ac, and p300 binding [57]. More recently, eRNAs were detected by native elongating transcript sequencing (NET-seq) [58], a method that captures the 3′ ends of nascent RNAs without the use of an *in vitro* nuclear run-on assay [58–60].

Further development of novel methods optimized for detection of eRNAs would be important to make enhancer identification feasible in a smaller sample size at cheaper cost with higher accuracy.

Recently, these functional genomics studies have been integrated into large-scale clinical genomics data and have brought valuable insights on the functional relevance of enhancers in many types of disease. In the following section we summarize and discuss the current understanding.

Genetic Variations within Enhancers Are Associated with Common Diseases

Similar to disease-causing variants in protein-coding genes, it is increasingly recognized that there is a tight association between common diseases and genetic variants at DNA regulatory regions including enhancers, where mutations might mediate transcriptional *cis* effects (reviewed in [1,2]).

Ernst *et al.* performed systematic mapping of enhancers using ChIP-seq of histone modifications and showed that disease-associated SNPs identified by GWASs are overrepresented within enhancers active in specific cell types [33]. Interestingly, GWAS SNPs are further shown to be

significantly enriched in super-enhancers [18] and stretch enhancers, which are similarly defined [61], but not in typical enhancers.

Recently, Farh *et al.* developed an algorithm to identify causal SNPs [referred to as probabilistic identification of causal SNPs (PICS)] for 21 autoimmune diseases and showed that a significant fraction of causal variants are located within enhancer containing regions determined by the H3K27ac ChIP-seq method in immune cells [62]. Many disease- or trait-associated SNPs were consistently overrepresented in the CAGE-defined enhancers (FANTOM5 enhancers) [41] (Figure 2A). Intriguingly, PICS SNPs were overrepresented within FANTOM5 enhancers to an even greater extent than protein-coding regions (Figure 2A). Further studies are needed to understand how variants in enhancers contribute to human diseases.

Genetic Mutations within Enhancers Are Associated with Tumors

Noncoding somatic mutations in cancer have been poorly explored, but several recent studies demonstrate links with tumorigenicity. Highly recurrent somatic mutations in the *telomerase reverse transcriptase* (*TERT*) promoter in melanomas were initially described [63]. Notably, whole-genome sequencing analyses of tumors highlighted new recurrent somatic mutations within enhancers [64–66]. More recently, Puente *et al.* showed that mutations within an enhancer reduced the expression of the B cell-specific TF paired box 5 (PAX5) in chronic lymphocytic leukemia [67].

Mansour *et al.* identified intergenic somatic mutations that created a binding motif for the MYB TF. This mutation generated a super-enhancer that led to overexpression of the *T cell acute lymphocytic leukemia 1* (*TAL1*) gene in a subset of patients with T cell acute lymphoblastic leukemia (T-ALL) [68]. Interestingly, in other subsets of patients overexpression of TAL1 has been shown to be mediated by a genomic fusion event occurring at the *TAL1* gene [69]. These data show the importance of examining enhancer regions of genes previously known to be causative. In addition, the 8q24 cancer risk variant rs6983267 [70,71] was shown to be positioned within an enhancer element, Myc-335 [72,73], that physically interacts with the MYC proto-oncogene promoter [73]. Sur *et al.* generated Myc-335 deletion mice, which strikingly displayed no apparent phenotype but were markedly resistant to intestinal tumorigenesis [74].

These findings prompted the examination of the enhancer region of genes shown to be mutated in the cancer genome project. In this study the Catalogue of Somatic Mutations in Cancer (COSMIC) [75] was searched for the corresponding cancer-associated GWAS SNP enhancer sequences defined in the FANTOM5 project. It was found that rs10411210, which is associated with colorectal cancer [76], is located on an enhancer linked to the *centrosomal protein 89kDa* (*CEP89*) gene, which is listed in the COSMIC database. Although rs10411210 was mapped to the *rhophilin*, *Rho GTPase binding protein 2* (*RHPN2*) gene based on the genomic distance, FANTOM5 CAGE data indicated that an enhancer covering rs10411210 targeted the *CEP89* gene in addition to the *RHPN2* gene based on the correlation of the transcription of the enhancer and that of target genes (Figure 2B).

Taking these findings together, genetic mutations that create or disrupt TF-binding sites could contribute to the pathogenesis of tumors by affecting enhancer activity and therefore should not be overlooked in making sense of cancer genomics data.

As a possible future direction, genetic examination of enhancers might contribute to the safe evaluation of emerging induced pluripotent stem (iPS) cell technology [77] (Box 3). In 2014, the first transplantation surgery using iPS cells was performed on a patient with age-related macular degeneration [78,79]. They first isolated the patient's dermal fibroblasts and generated iPS cells.

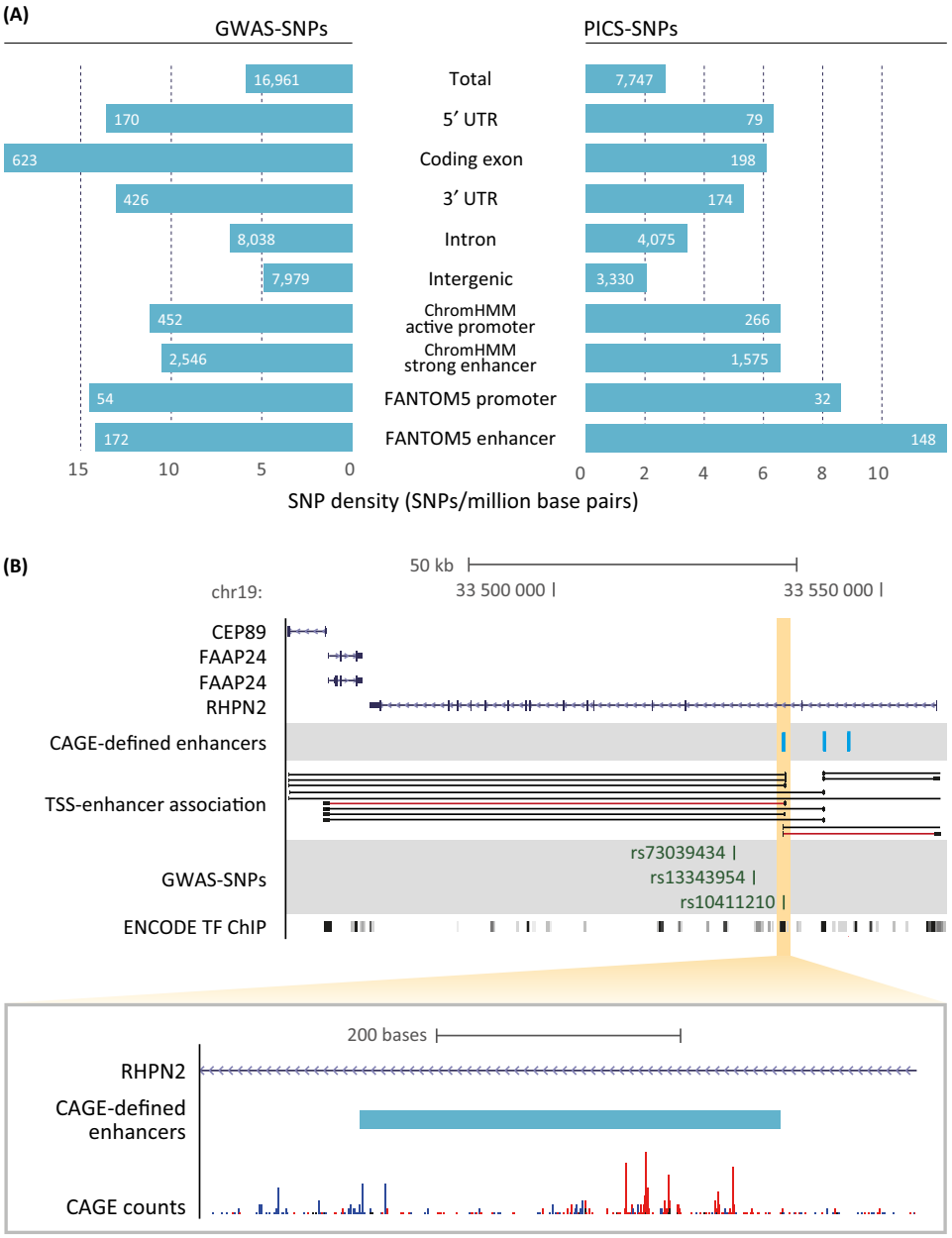


Figure 2. Genetic Alterations within Enhancers Associated with Diseases. (A) Distribution of genome-wide association study (GWAS) SNPs and probabilistic identification of causal SNPs (PICS) SNPs (causal SNPs for 21 autoimmune diseases) over distinct genomic regions. ChromHMM promoters and enhancers are characterized by a combinatorial pattern of histone modifications [33] and FANTOM5 promoters and enhancers are identified by cap analysis of gene expression (CAGE) [41,99]. The number of SNPs in each region is shown next to each bar. (B) UCSC Genome browser view of rs10411210 SNP associated with colorectal cancer. rs10411210, which is associated with colorectal cancer [76], is located within an enhancer linked with the *CEP89* gene, which is listed in the COSMIC database [75]. Although rs10411210 was mapped to the *RHPN2* gene based on the genomic distance, the FANTOM5 CAGE data indicate that this enhancer targets the *CEP89* gene in addition to the *RHPN2* gene (as highlighted in red) based on the correlation of enhancer and promoter transcriptions. A detailed view is shown at the bottom.

Box 3. iPS Cell Technology

iPS cells are a type of pluripotent stem cell. iPS cells are generated by direct reprogramming of mature somatic cells through the introduction of genes encoding key TFs such as Oct4, Sox2, cMyc, and Klf4 [77]. Therefore, iPS cells can be obtained without manipulation of the preimplantation-stage embryo, solving the most worrisome ethical issue of embryonic stem cells (ESCs). Furthermore, tailor-made iPS cells prepared from patients themselves are unlikely to trigger an immune rejection reaction [100]. Therefore, iPS cell technology is expected to have enormous potential, especially in the field of regenerative medicine.

The iPS cells were differentiated into retinal pigment epithelial cells and then grown into a sheet for implantation. Regenerative medicine utilizing iPS cells is being pursued for the treatment of many other diseases (reviewed in [80]). Seen as a major concern is the safety issue with respect to genomic alterations and tumorigenicity [81–83]. Here we propose that mutations within enhancers, in addition to protein-coding sequences, should be carefully examined in genomics studies in light of the significant association with tumorigenicity.

Key Figure

Schematic Representation of a Transcribed Enhancer and Comparison of Enhancer Identification Methods

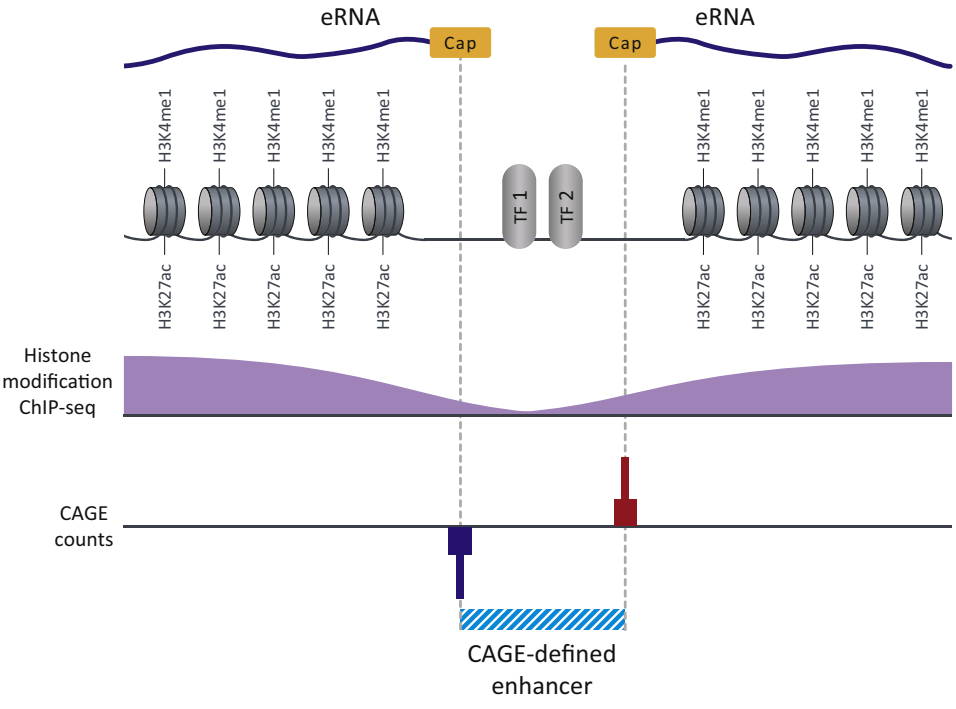


Figure 3. Transcription factors (TFs) and coactivators occupy the active enhancer region. In addition, the active enhancer initiates RNA polymerase II (RNAPII) to generate bidirectional 5′-capped enhancer RNA (eRNA). Nucleosomes flanking the active enhancer region are marked by histone H3 lysine 4 monomethylation (H3K4me1) and H3K27 acetylation (H3K27ac). Shown below is a genome browser view of representative ChIP-seq of histone modification and cap analysis of gene expression (CAGE) data. CAGE detects the bidirectionally transcribed eRNAs at nucleotide resolution whereas ChIP-seq of H3K4me1 and H3K27ac shows broad peaks surrounding the enhancer region.

Concluding Remarks

Over 30 years after the first identification of enhancers, the advent of next-generation sequencing has changed the paradigm. Various experimental and computational methods were developed to identify enhancers on a genome-wide scale, providing deeper understanding of the functional and structural characteristics of enhancers. In addition, international consortiums such as ENCODE and FANTOM have provided valuable large-scale resources for further investigations[ii,vi]. However, various methods predict enhancers by capturing different molecular features associated with enhancers (e.g., NDRs, TF binding, histone modifications) and the term enhancer is used ambiguously, with increasing number of methods to identify them. For instance, ChIP-seq identifies 'enhancers' by detecting histone modifications occurring in broad regions surrounding the NDRs that are bound by TFs, which results in limited base resolution. By contrast, CAGE identifies transcribed enhancers at exceedingly high nucleotide resolution by detecting the bidirectionally transcribed eRNAs, although detection of these eRNAs requires high sequencing coverage due to their low level of expression (Figure 3, Key Figure) (see Outstanding Questions). Importantly, each method has its advantages and disadvantages, as discussed here. Therefore, a combination of complementary methods, rather than relying on just a single method, might provide better enhancer prediction.

We have an increasingly clearer picture of genetic alterations within enhancers contributing to diseases. Intriguingly, disease-associated SNPs were overrepresented within CAGE-defined enhancers to an even greater extent than protein-coding regions. Furthermore, recent studies demonstrate that mutations within enhancers alter the enhancers' activities in several cancers. These findings highlight the importance of enhancer identification and emphasize its potential direct roles in various clinical settings including GWASs, cancer genomics studies, and safety evaluations for iPS cell-based regenerative medicine.

Acknowledgments

The authors apologize to those whose work is not cited owing to space limitations. They thank Mr Jayson Harshbarger for critical reading and editing of the manuscript. Work in the laboratory of Y.H. is supported by grants from the Japanese Ministry of Education, Culture, Sports, Science, and Technology (MEXT).

Resources

[i] VISTA Enhancer Browser: http://enhancer.lbl.gov/

[ii] ENCODE Consortium: http://www.genome.gov/encode/

[iii] Factorbook: http://factorbook.org/

[iv] UCSC Genome Browser: http://genome.ucsc.edu/

[v] ENCODE downloads: http://genome.ucsc.edu/ENCODE/downloads.html

[vi] FANTOM Consortium: http://fantom.gsc.riken.jp/

[vii] PrESSTo Human Enhancers: http://enhancer.binf.ku.dk/enhancers.php

References

1. Epstein, D.J. (2009) *Cis*-regulatory mutations in human disease. *Brief. Funct. Genomic. Proteomic.* 8, 310–316
2. Mathelier, A. *et al.* (2015) Identification of altered *cis*-regulatory elements in human disease. *Trends Genet.* 31, 67–76
3. Schwanhausser, B. *et al.* (2011) Global quantification of mammalian gene expression control. *Nature* 473, 337–342
4. Juven-Gershon, T. and Kadonaga, J.T. (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.* 339, 225–229
5. Ong, C.T. and Corces, V.G. (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* 12, 283–293
6. Shlyueva, D. *et al.* (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15, 272–286
7. Banerji, J. *et al.* (1981) Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27, 299–308
8. Moreau, P. *et al.* (1981) The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res.* 9, 6047–6068
9. Banerji, J. *et al.* (1983) A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* 33, 729–740
10. Gillies, S.D. *et al.* (1983) A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* 33, 717–728
11. Mercola, M. *et al.* (1983) Transcriptional enhancer elements in the mouse immunoglobulin heavy chain locus. *Science* 221, 663–665
12. Spitz, F. and Furlong, E.E. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* 13, 613–626

Outstanding Questions

What is the function of eRNAs? Several reports showed that knockdown of eRNAs reduced the expression of target mRNAs [43,44]. However, the detailed molecular mechanism of eRNAs remains elusive. To what extent does each eRNA contribute to enhancer function, given that a fraction of untranscribed enhancers exhibit enhancer activity in a reporter assay and a fraction of transcribed enhancers do not [41]? However, this might be explained by false-negative or -positive identification of transcribed enhancers.

Some eRNAs are unidirectionally transcribed, polyadenylated, or uncapped, showing heterogeneity. What causes these differences?

Various approaches have been developed to predict enhancers on a genome-wide scale. However, there is no consensus on which method is the most reliable. Each method has its advantages and disadvantages. We argue that eRNA-based methods can determine transcribed enhancers at high nucleotide resolution, but eRNAs could be readily detected only by high sequencing coverage due to their low level of expression. It is also sometimes difficult to distinguish promoters and enhancers because they have many similarities [49]. Furthermore, the term enhancer is used ambiguously, with an increasing number of methods that capture different molecular features of enhancers such as histone modifications and eRNA expression. Therefore, what is the best method to determine enhancer elements and what is the fundamental element of enhancers?

Recurrent cancer somatic mutations and disease-associated genomic variants are identified within enhancers. However, how genomic alterations within enhancers lead to disease is not fully understood. The phenotypic impact of mutations within enhancers might be studied by recent genome editing tools such as CRISPR/Cas9. Elucidation of how enhancers work in development or pathology might help in the design of therapeutic approaches such as chemical inhibition of disease-specific eRNAs.

13. Buecker, C. and Wysocka, J. (2012) Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet.* 28, 276–284
14. Kagey, M.H. *et al.* (2010) Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 467, 430–435
15. Thurman, R.E. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature* 489, 75–82
16. Zhou, V.W. *et al.* (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.* 12, 7–18
17. Calo, E. and Wysocka, J. (2013) Modification of enhancer chromatin: what, how, and why? *Mol. Cell* 49, 825–837
18. Hnisz, D. *et al.* (2013) Super-enhancers in the control of cell identity and disease. *Cell* 155, 934–947
19. Loven, J. *et al.* (2013) Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* 153, 320–334
20. Whyte, W.A. *et al.* (2013) Master transcription factors and Mediator establish super-enhancers at key cell identity genes. *Cell* 153, 307–319
21. Pott, S. and Lieb, J.D. (2015) What are super-enhancers? *Nat. Genet.* 47, 8–12
22. Woolfe, A. *et al.* (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* 3, e7
23. Pennacchio, L.A. *et al.* (2006) *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* 444, 499–502
24. Visel, A. *et al.* (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.* 40, 158–160
25. Visel, A. *et al.* (2007) Enhancer identification through comparative genomics. *Semin. Cell Dev. Biol.* 18, 140–152
26. Hardison, R.C. and Taylor, J. (2012) Genomic approaches towards finding *cis*-regulatory modules in animals. *Nat. Rev. Genet.* 13, 469–483
27. Visel, A. *et al.* (2007) VISTA Enhancer Browser – a database of tissue-specific human enhancers. *Nucleic Acids Res.* 35, D88–D92
28. Visel, A. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457, 854–858
29. Buenrostro, J.D. *et al.* (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin DNA-binding proteins and nucleosome position. *Nat. Methods* 10, 1213–1218
30. Tsompana, M. and Buck, M.J. (2014) Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* 7, 33
31. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74
32. Wang, J. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* 22, 1798–1812
33. Ernst, J. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43–49
34. Zentner, G.E. and Henikoff, S. (2012) Surveying the epigenomic landscape, one base at a time. *Genome Biol.* 13, 250
35. Dogan, N. *et al.* (2015) Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics Chromatin* 8, 16
36. Kim, T.K. *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465, 182–187
37. Tuan, D. *et al.* (1992) Transcription of the hypersensitive site HS2 enhancer in erythroid cells. *Proc. Natl. Acad. Sci. U.S.A.* 89, 11219–11223
38. Lam, M.T. *et al.* (2014) Enhancer RNAs and regulated transcriptional programs. *Trends Biochem. Sci.* 39, 170–182
39. Koch, F. *et al.* (2011) Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat. Struct. Mol. Biol.* 18, 956–963
40. Djebali, S. *et al.* (2012) Landscape of transcription in human cells. *Nature* 489, 101–108
41. Andersson, R. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461
42. Pefanis, E. *et al.* (2015) RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. *Cell* 161, 774–789
43. Melo, C.A. *et al.* (2013) eRNAs are required for p53-dependent enhancer activity and gene transcription. *Mol. Cell* 49, 524–535
44. Li, W. *et al.* (2013) Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* 498, 516–520
45. Mousavi, K. *et al.* (2013) eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol. Cell* 51, 606–617
46. Schaukowitch, K. *et al.* (2014) Enhancer RNA facilitates NELF release from immediate early genes. *Mol. Cell* 56, 29–42
47. Pnueli, L. *et al.* (2015) RNA transcribed from a distal enhancer is required for activating the chromatin at the promoter of the gonadotropin alpha-subunit gene. *Proc. Natl. Acad. Sci. U.S.A.* 112, 4369–4374
48. Shiraki, T. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15776–15781
49. Andersson, R. *et al.* (2015) A unified architecture of transcriptional regulatory elements. *Trends Genet.* 31, 426–433
50. Yao, P. *et al.* (2015) Coexpression networks identify brain region-specific enhancer RNAs in the human brain. *Nat. Neurosci.* 18, 1168–1174
51. Yamashita, R. *et al.* (2011) Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res.* 21, 775–789
52. Ni, T. *et al.* (2010) A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat. Methods* 7, 521–527
53. Melgar, M.F. *et al.* (2011) Discovery of active enhancers through bidirectional expression of short transcripts. *Genome Biol.* 12, R113
54. Core, L.J. *et al.* (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 322, 1845–1848
55. Kwak, H. *et al.* (2013) Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* 339, 950–953
56. Nechaev, S. *et al.* (2010) Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science* 327, 335–338
57. Scruggs, B.S. *et al.* (2015) Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol. Cell* 58, 1101–1112
58. Mayer, A. *et al.* (2015) Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* 161, 541–554
59. Churchman, L.S. and Weissman, J.S. (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* 469, 368–373
60. Nojima, T. *et al.* (2015) Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* 161, 526–540
61. Parker, S.C. *et al.* (2013) Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U.S.A.* 110, 17921–17926
62. Farh, K.K. *et al.* (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518, 337–343
63. Huang, F.W. *et al.* (2013) Highly recurrent TERT promoter mutations in human melanoma. *Science* 339, 957–959
64. Fredriksson, N.J. *et al.* (2014) Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* 46, 1258–1263
65. Weinhold, N. *et al.* (2014) Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* 46, 1160–1165
66. Melton, C. *et al.* (2015) Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.* 47, 710–716
67. Puente, X.S. *et al.* (2015) Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* 526, 519–524

68. Mansour, M.R. *et al.* (2014) Oncogene regulation An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* 346, 1373–1377
69. Brown, L. *et al.* (1990) Site-specific recombination of the *tal-1* gene is a common occurrence in human T cell leukemia. *EMBO J.* 9, 3343–3351
70. Tomlinson, I. *et al.* (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat. Genet.* 39, 984–988
71. Yeager, M. *et al.* (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* 39, 645–649
72. Tuupanen, S. *et al.* (2009) The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat. Genet.* 41, 885–890
73. Pomerantz, M.M. *et al.* (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.* 41, 882–884
74. Sur, I.K. *et al.* (2012) Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* 338, 1360–1363
75. Forbes, S.A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* 39, D945–D950
76. Houlston, R.S. *et al.* (2008) Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.* 40, 1426–1435
77. Takahashi, K. *et al.* (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131, 861–872
78. Kamao, H. *et al.* (2014) Characterization of human induced pluripotent stem cell-derived retinal pigment epithelium cell sheets aiming for clinical application. *Stem Cell Rep.* 2, 205–218
79. Cyranoski, D. (2014) Japanese woman is first recipient of next-generation stem cells. *Nature* Published online September 12, 2014. http://dx.doi.org/10.1038/nature.2014.15915
80. Okano, H. and Yamanaka, S. (2014) iPS cell technologies: significance and applications to CNS regeneration and disease. *Mol. Brain* 7, 22
81. Mayshar, Y. *et al.* (2010) Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. *Cell Stem Cell* 7, 521–531
82. Gore, A. *et al.* (2011) Somatic coding mutations in human induced pluripotent stem cells. *Nature* 471, 63–67
83. Sugiura, M. *et al.* (2014) Induced pluripotent stem cell generation-associated point mutations arise during the initial stages of the conversion of these cells. *Stem Cell Rep.* 2, 52–63
84. Wasserman, W.W. and Fickett, J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* 278, 167–181
85. Chen, X. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133, 1106–1117
86. Zinzen, R.P. *et al.* (2009) Combinatorial binding predicts spatio-temporal *cis*-regulatory activity. *Nature* 462, 65–70
87. May, D. *et al.* (2012) Large-scale discovery of enhancers from human heart tissue. *Nat. Genet.* 44, 89–93
88. Dorschner, M.O. *et al.* (2004) High-throughput localization of functional elements by quantitative chromatin profiling. *Nat. Methods* 1, 219–225
89. Giresi, P.G. *et al.* (2007) FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res.* 17, 877–885
90. Heintzman, N.D. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39, 311–318
91. van Steensel, B. and Dekker, J. (2010) Genomics tools for unraveling chromosome architecture. *Nat. Biotechnol.* 28, 1089–1095
92. Li, G. *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148, 84–98
93. Melnikov, A. *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* 30, 271–277
94. Arnold, C.D. *et al.* (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339, 1074–1077
95. Lai, F. *et al.* (2015) Integrator mediates the biogenesis of enhancer RNAs. *Nature* 525, 399–403
96. Rosenbloom, K.R. *et al.* (2013) ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.* 41, D56–D63
97. de Hoon, M. *et al.* (2015) Paradigm shifts in genomics through the FANTOM projects. *Mamm. Genome* 26, 391–402
98. Arner, E. *et al.* (2015) Gene regulation Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* 347, 1010–1014
99. The FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* (2014) A promoter-level mammalian expression atlas. *Nature* 507, 462–470
100. Zhao, T. *et al.* (2011) Immunogenicity of induced pluripotent stem cells. *Nature* 474, 212–215