The intriguing complexities of mammalian gene regulation: how to link enhancers to regulated genes. Are we there yet?

Bence Daniel[1], Gergely Nagy[1] and Laszlo Nagy[1,2,3]

[1] Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Debrecen, Egyetem tér 1. Debrecen, Hungary H-4010

[2] MTA-DE „Lendület” Immunogenomics Research Group, University of Debrecen, Egyetem tér 1. Debrecen, Hungary

[3] Sanford-Burnham Medical Research Institute, 6400 Sanger Road, Orlando, FL USA 32827

Abstract

The information encoded in genomes supports the differentiation and function of the more than 200 unique cell types, which exist in various mammalian species. The major mechanism driving cellular differentiation and specification is differential gene expression regulation. Cis-acting enhancers and silencers appear to have key roles in regulating the expression of mammalian genes. However these cis-acting elements are often located very far away from the regulated gene. Therefore it is hard to find all of them and link them to the regulated gene. An intriguing and unresolved issue of the field is to identify all of the enhancers of a particular gene and link these short regulatory sequences to the genes they regulate and thus reliably identify gene regulatory enhancer networks. Recent advances in molecular biological methods coupled to Next-Generation Sequencing (NGS) technologies have opened up new possibilities in this area of genomics. In this review we summarize the technological advances, bioinformatics challenges and the potential molecular mechanisms allowing the construction of enhancer networks operating in specific cell types and/or activated by various signals.

Keywords

Gene expression regulation

Enhancers

Genomics

Transcriptome

Cis-regulatory elements

Chromosome conformation

Why mammalian gene regulation is so hard to study?

Gene expression is the result of a very complex process achieved by the coordinated action of multiple layers of regulators. In prokaryotes, genes are organized into operons and using a single promoter the entire stretch of DNA is transcribed into RNA (Figure 1A). These so-called “polycistronic” RNAs often encode functionally related members of an enzyme cascade regulating a particular metabolic process. Thus, in most of the cases prokaryotes utilize one promoter-proximal, restricted cis-element to initiate transcription. However, eukaryotic organisms have evolved to utilize much more complex mechanisms to regulate gene expression. In fact, one of the driving forces of eukaryotic evolution is believed to be the introduction of elaborate gene regulatory circuits. This is, in part, manifested in the concept of C value enigma, which is the observation that genome size does not correlate with organismal complexity [1]. The number of protein coding genes also does not show correlation with complexity.

Unlike prokaryote genes, eukaryotic ones are “monocistronic” and their regulation is usually much more complex. The fact that genes can have multiple promoters with unique promoter elements makes the picture even more crowded and complicated. In addition, probably most if not all eukaryotic genes possess intergenic as well as intragenic cis-regulatory elements (enhancers/silencers) to fine-tune their expression in a cell type and/or biological context dependent manner (Figure 1B, 2B). To clearly understand the detailed molecular mechanisms controlling gene expression, one needs to identify the factors responsible for gene regulation and their precise action on one or multiple well-defined cis-regulatory element(s). Before the genomic era, investigations were limited methodologically and gene regulation was studied with methods restricted to the analysis of the expression of a few genes and/or biased/restricted to the immediate vicinity of a given gene. The classical way of studying gene expression used a set of so-called “promoter bashing approaches” evaluating the genomic regions in the close proximity of the transcription start site (TSS) to identify the core sequence driving the expression of the given gene [2]. These approaches were based on transient transfections and deletion and insertion mutagenesis and are still used to date. The discovery of enhancers predicted the complex regulation of a gene, because these cis-regulatory elements can be located, at least in principle, far away and either upstream 5’ or downstream 3’ of the regulated gene [3]. The inherent pitfalls of the first studies were recognized soon, but without the technical advances no one could easily go beyond the technical limitations to dissect the regulation of gene networks, not even just a particular gene. Therefore most studies identified promoter proximal elements usually restricted to 10-20 kb upstream of the designated promoter. Using these typical "promoter bashing" technologies researchers were not able to consider and/or evaluate the contribution of multiple and/or far way enhancers or intra- or interchromosomal interactions. This represented bias and created a roadblock in understanding complex gene regulation.

An additional important aspect of gene expression regulation research is that one would need methods to assess the expression of several dozens of genes at the same time in order to determine the impact of a particular signal on what is now called global gene expression.

The first attempt to provide information regarding global gene expression was reported in 1982, when the first gene array was constructed to assay 378 transcripts in normal and tumor tissues [4]. Until 1995 several studies were published using this filter paper spotted technology to document the gene expression pattern of cancerous versus normal cells and also the anti-proliferative action of interferons [5], [6]. The usage of miniaturized microarray technology was first applied to measure the differential expression of 45 Arabidopsis genes [7]. These initial efforts launched the nowadays' well-known microarray technology which transformed the field of transcription regulation.

New NGS-driven enabling technologies have emerged

As the result of the genome programs and the technological revolution coming with it, major breakthroughs helped scientists working on the field of mammalian gene expression regulation.

As a key example the profiling of transcription factor binding sites (TFBSs) and histone modification patterns along with nucleosome positions have been determined by linking Chromatin Immunoprecipitation (ChIP) to microarray technology for a review see [8]. This allowed the identification of histone modification and /or TFBSs in a quasi genome-wide fashion and showed how old molecular biology techniques (i.e. immunprecipitation) can be effectively coupled to genome-based approaches (i.e.

microarrays of genomic DNA fragments). During the last several years, sequencing technologies have emerged and almost fully replaced the microarray-based methods though. With the ability to sequence tens of millions of short DNA sequences in a parallel fashion, more applications, which could be only imagined before, became reality. This rapid evolution of sequencing, now termed Next-Generation Sequencing (NGS), entirely transformed the field of gene expression research, along with other areas of research. The combination of NGS with simple molecular biology approaches proved to be very effective and led to the rapid generation of genome-wide data in a number of areas including whole genome sequencing, mRNA sequencing and revealing MNase and DNase I hypersensitive sites as well as reviewed in [8] (Table 1). For high throughput TSS sequencing, similar methods were developed simultaneously for different sequencing platforms based on the CAGE (Cap Analysis of Gene Expression), 5' SAGE (Serial Analysis of Gene Expression) and 5' RACE (Rapid Amplification of cDNA Ends) methods [9]. ChIP has been one of the earliest applications linked to NGS and led to the determination of histone modification patterns, co-factor and transcription factor binding sites throughout the genome [8]. These results supplied the first set of evidence that cis-acting elements are likely to be widespread in the mammalian genomes and are mostly located in intergenic regions also at large distances from the TSSs of the putative targeted gene.

Our ability to detect intra- and interchromosomal interactions has also changed quite a bit. The development of technologies such as 3C greatly increased one's options to reveal the interaction map of *cis*-elements beyond their interactions with the regulated genes [10]. The 3C method is based on the fixation of chromatin loops mediated by protein complexes bound to DNA. After stabilizing the interactions a restriction enzyme is used to cut the genome into smaller pieces containing the compatible sticky ends. Then the chromatin is subjected to ligation in a highly diluted fashion. This reaction favors intramolecular ligation events, thus capable of connecting those genomic regions residing in the proximity of each other.

Over the last decade 3C has been linked to NGS and also combined with ChIP resulted in the following technologies (3C-Seq, 4C-Seq, 5C, ChIA-PET, Hi-C) (Table 2). Depending on the biological question, one can choose these methods to reveal the physical conformation of the genome or a particular locus. The 4C assay can be very useful if one genomic element and its interactions are examined. Initially, libraries obtained from 4C experiments were hybridized to microarrays to get insights into chromatin interactions [11]. 3C-sequencing basically holds the same advantage as the 4C experiments and it can be used for detecting one to all interactions [12]. 5C overcomes the disadvantage of 4C and 3C-sequencing and offers the opportunity to map all the interactions in a large chromosomal territory [13]. 4C [14] and 5C [15] has been linked to NGS now providing genome-wide information. However Hi-C (Table 2) has been the most powerful among these by having the ability to map all the interactions genome-wide in an unbiased way [16]. ChIA-PET [17] is the genome-wide version of the method called ChIP-loop [18] in which the combination of ChIP and 3C is utilized to map the interactions between any two loci bound by the investigated protein (Table 2).

These technological developments apparently produced the possibility to reveal interaction maps about the identified cis-acting elements and their promoters. In addition, these can inform one about the organization of higher-order chromatin structure in the nucleus. These technologies are quite robust, provide a lot of data, prone to artifacts and require significant bioinformatics efforts to analyze and interpret the data.

The need for bioinformatics

The necessity of bioinformatics became apparent with the delivery of sequencing data to the molecular biology laboratory. Post-genomic bioinformatics has been evolving in concert with the emergence of NGS methods, but such evolution is not without difficulties and dead ends especially considering the fact that two different fields, molecular biology and informatics needed to merge. It is safe to say that as of today there are still no standardized algorithms, or even consensus quality control metrics to qualify and interpret NGS data. The fact that the data sets need to be deposited in central repositories is a necessary step and allows re-analyses by independent groups and/or independent methods. The largest such source is the NCBI Sequence Read Archive, which contains the raw sequence reads of several tens of thousands of NGS samples. These sequences can be freely downloaded and reanalyzed, mapped to the appropriate reference genome. Alignment of spliced reads for RNA-Seq was a unique challenge which had been overcome e.g. with the TopHat software [19], which finally led to a now widely used RNA-Seq analysis pipeline.

Determination of cistromic and epigenomic enrichments are accomplished with several distinct algorithms specialized for shorter or broader regions with different, e.g. peak-like, column-like, extended or valley-like shapes. Differences in read distributions of a sample lead to slightly or sometimes not so slightly different results in terms of number of the predicted binding sites overlaps between occupied genomic regions, etc. There is also no real consensus in the way the results are plotted and presented in publications. These can lead to confusions and disagreements regarding the cistromes of transcription factors or the effects of various signals.

Genomic ChIP-ed fragments of transcription factors ideally show a Gaussian distribution near the TFBSs forming peaks after the alignment. The initial studies used SOAP, Bowtie and BWA, but there are other sequence read mapper tools as well for the alignment. There are additional peak caller programs have been developed using different algorithms: CSAR searches for read enrichments following Poisson distribution. MACS and Homer work based on the local density of the reads with Poisson distribution. CisGenome uses negative binomial, and ZINBA uses zero-inflated negative binomial distribution. There are more sophisticated methods for peak calling such as the BayesPeak using the Hidden Markov Model, see review about peak caller tools [20]. This list of approaches testifies that there is still a lack of consensus on what to call a peak and how to count it.

It appears though that peak caller tools work relatively well depending on the quality of the sample libraries and give a quality score, the edges and the summit of the peaks. Peak width is usually a technical issue but the summit has a biological relevance as it is very close to the putative TFBS(s). The first widely used package, which was able to search for motif enrichments and to map the found matrices (back) to sequences thus designating the TFBSs was the MEME-MAST toolkit [21]. Based on the validated TFBSs more motif matrix databases were established: TRANSFAC [22], MEME and JASPAR [23] have databases with their own similarity weight matrix formats, which are used in several other databases as well. Homer also developed a motif enrichment and TFBS searching system, which includes others’ matrices as well as the ones with Homer’s own format enriched from different ChIP-Seq samples

[24]. There is a special ChIP-Seq method called ChIP-exo, developed directly for the detection of TFBSs at a single nucleotide resolution. Fragments are shortened by an exonuclease from the 5’ up to the TFBS, thus resulting in a column-like shape after alignment marking the exact place of the DNA binding [25]. Application of this method might result in a much better resolution in determining TFBSs.

Histone modifications usually cover broader regions of the genome, which calls for different kinds of algorithms. SICER [26] and ZINBA were developed for this purpose but certain peak callers are also able to find these kinds of regions by using different parameters, such as MACS2 and Homer. DNA binding of transcription factors disrupts histone continuity establishing the so-called nucleosome-free (or more precisely nucleosome-depleted) regions (NFRs) with valley-like shapes in the histone (modification) landscapes. There were several methods allowing the detection of nucleosome occupied and depleted regions. The first one is based on ChIP-chip [27], [28], whilst others are based on MNase or ChIP-Seq data [29]. Homer includes an NFR prediction function for ChIP-Seq data as well, and we have also developed a method to find putative regulatory regions based on active histone mark landscapes [30]. Special NGS methods emerged in order to determine the regulatory regions and as FAIRE- (Formaldehyde-Assisted Isolation of Regulatory Elements) [31] and Sono-Seq [32] give peak-like enrichments, peak callers are also suitable in these cases to determine these NFRs as well. Sono-Seq is a simple method to find accessible chromatin regions using a size selection following the sonication of the cross-linked chromatin. More recently, predictors of broader regions (SICER, ZINBA) and NFRs are becoming suitable for the processing of DNase I, MNase-Seq or other kinds of NGS data.

How to link the regulator to the regulated?

The combined usage of molecular, NGS and bioinformatics methods and approaches continue to provide useful and a very large amount of information about the location of cis-acting elements and the gene expression profile of a given cell type. However, aligning the regulatory elements to the affected genes remained very cumbersome if not impossible. Linking TFBSs simply to the closest regulated gene is not so reliable in case of larger distances between the proposed enhancers and the regulated genes, without evidence for these interactions unless these are validated by methods such as Chromosome Conformation Capture (3C) [10]. 3C, if done quantitatively, is suitable for measuring the interaction frequency between any two loci in the genome, however prior information (e.g. TFBSs or histone modification patterns) of the given loci are critical for the experimental design. The development of Hi-C overcomes this problem [16]. This method is based on 3C, of which ligation products are sequenced on an NGS platform documenting all genome-wide interactions. However its resolution is far from ideal to map enhancer-promoter interactions. Fullwood et al. reported a new technology called chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) in 2009 (Table 2). The method itself is the combination of ChIP and 3C, and first applied to map the chromatin interactions influenced by the estrogen receptor alpha in the human genome [17]. The usage of this technology provided a tool to understand how a transcription factor can act from long distances, and implied that chromatin interactions are one of the driving molecular mechanisms for regulating gene expression in the mammalian genomes.

Development of Global Run-On sequencing (GRO-Seq) was an additional breakthrough. The procedure is based on the classical nuclear run-on transcription assay, which has been used in molecular biology for decades, providing a snapshot about the level of *in vivo* (at least in an intact nucleus) synthesized nascent RNAs [33]. If done as a time course, the dynamics of the induced/changing transcription can be assessed and even quantitated. GRO-Seq was first applied to primary human lung fibroblasts in order to map the amount, position and orientation of the transcriptionally engaged RNA polymerases [34]. The obtained results were very striking, showing that about 30% of the human genes are occupied by active polymerases, genes are transcribed beyond the 3' end of the annotated regions, and surprisingly, most promoters possess engaged polymerases in the opposite orientation to the annotated gene. This divergent transcription is associated with active genes but usually is not elongated efficiently to the upstream regions relative to the TSS [34].

The technical advance made it possible to go forward using these approaches (Table 1.) to reveal the complex regulation of the genes in a genome-wide manner, however the combination of the approaches is necessary to understand and to be able to explain the most exciting findings. In the following part of this review we make an attempt to summarize the recent developments in the field of data integration based genomics, regarding cistromic, transcriptomic and chromatin structure related studies, which are based on molecular biology of NGS coupled technologies.

Enhancers, the main drivers of gene expression

In complex genomes *cis*-acting elements are dispersed and can be located over several hundred kilobases far from their targeted genes [35]. Developments regarding NGS rapidly expanded our view about the putative location of regulatory elements and shed light on the problematic nature of assigning enhancers to genes. However it has been shown in one of the earliest studies that their proper action is indispensable and mutation occurring within the core enhancer sequence may result in congenital disease [36]. Based on these, the proper annotation of enhancers has become an important and widely studied issue in the field of transcriptional regulation and an absolute requirement for proper annotation of genome function. Enhancers are DNA sequences with the ability to recruit various types of transcription factors for the interaction with the mediator complex as well as with the members of the (pre)-initiation complex. By looping mechanisms the complexes assembled on the DNA can facilitate RNA-polymerase II (RNAPII) binding to the promoter, thus the initiation of gene transcription [37]. It has been also shown that transcription factors bound to enhancers are able to recruit ATP-dependent chromatin remodeling enzyme complexes affecting the chromatin structure to establish NFRs, thus facilitating transcription factor binding, transcription initiation and elongation [38].

The appearance of MNase-, DNase I-, and ChIP-Seq (Table 2) as main tools to interrogate open chromatin, transcription factor binding and epigenetically marked histone landscapes has greatly improved our understanding about the main features of enhancers. Results from genome-wide studies mapping nucleosome occupancy indicate that at cis-regulatory elements, histone replacement is more enhanced than at other genomic locations not harboring such enhancer-like properties [39]. Active promoters/TSSs are hardly occupied by nucleosomes as well, thus these form also NFRs. These results suggested that nucleosome stability contributes to gene regulation [40]. Later it has been shown that the two alternative, minor histone variants, H3.3 and H2A.Z are enriched near NFRs [41].

High-resolution co-activator: CREB-binding protein (CBP) and p300, ChIP-Seq enrichments provided further insights into the chromatin signatures of enhancers. These proteins interact with various transcription factors and possess histone acetyl-transferase activity, which makes them capable to modify histones [42]. Several studies showed that these factors are good predictors of enhancer function in a tissue specific manner as reviewed in [43]. These results suggest that CBP/p300 co-factors are key functional components of the enhancer binding complexes.

Certain histone modifications also participate in cis-element function and it has been shown that the main characteristic features of active promoters are defined by the residence of RNAPII and TBP-associated factor 1 (TAF1), marked by NFRs flanking with trimethylated histone H3 at lysine 4 (H3K4me3), acetylated H3 (H3ac) and TFIID [44]. On the other hand, as previously mentioned p300 is one of the most well-documented active enhancer marks along with the enriched H3K4me1, H3K4me2 and H3K27ac for a review see [45]. These observations were confirmed in different cellular model systems leading to the identification of enhancer repertoires of a given cell type [46], [47]. Based on these studies, the identification of enhancers and their characteristic features are relatively straightforward, however the annotation process to the affected genes remained largely elusive.

Utilizing the features of RNA-Seq and GRO-Seq, another useful feature has been described, namely that transcription can be detected at active enhancers [48] [49]. The observation added another layer to enhancer features and it turned out that this might be the most reliable indicator that the enhancer actively participates in gene regulation [50]. We will elaborate on this later.

Linking enhancers to regulated genes

As far as the genomic features of enhancers are recognized, their identification has been more efficient using enhancer prediction methods based on both evolutionary conservation [51], and ChIP-Seq results [52]. Based on factor occupancy and histone marks the enhancers can be identified, but their targeted genes are harder to find. In the absence of a better method, most studies in the field used simply proximity based predictive approaches to assign enhancers to the regulated genes [53], [54]. Over the past 12 years the development of 3C methodology and its combination with NGS technology rapidly changed our view about genome structure though [55]. Traditionally, nuclear organization was evaluated by microscopy based methods. Since then, different 3C related methods have been evolved leading to the 3D determination of chromatin structure at various gene loci [55]. The advantage of these 3C methods is their higher resolution compared to microscopy and allowing to analyze a single gene and its interaction profile. Importantly, 3C-based approaches will always need a helping hand from the side of microscopy to fully uncover the shape of the genome and to obtain the most reliable interactions [55] (see also later).

The simplest example is the comparison of gene expression in active and inactive chromatin regions. Chromatin segregation into active and inactive regions raised the question whether positioning into these regions affects gene expression. This was clearly documented using fluorescent *in situ* hybridization (FISH) that certain regions change their nuclear position upon the activating stimulus [56]. Probably some of these changes at the level of gene expression are attributable to *cis*-acting elements such as enhancers. The correlation between nuclear position and gene expression has been shown in several studies, for a review see [55]. Silent genes are localized closer to nuclear lamina than their active counterparts and supported the basis of the so-called „position effect", which describes the behavior and action of *cis*-regulatory elements in the context of higher-order chromatin structure [55]. Another emerging concept is the existence of subnuclear compartments enriched in transcription, so called transcriptional factories [56]. These would be highly relevant for enhancer activity and might even suggest that enhancers and promoters co-localize in such subnuclear regions.

ENCODEing the functional elements

Humans are estimated to have about 20,000 protein-coding genes, but this covers only 1.5% of the entire genome. The other 98% accounts for intronic and intergenic regions, non-coding RNAs and short or long interspersed elements [57]. ENCODE is a project launched in 2003 aimed at identifying all the *cis*-acting elements in the human genome via the collaboration of several research groups as part of this specialized consortium [58]. This initiative is the continuation of the Human Genome Project, but in this particular case the final aim was the identification of functional DNA sequences that act at the protein and RNA levels in a given cell type [59]. The functional DNA sequences defined as distinct genomic sections encoding a specified output for instance, protein product or non-coding RNA. Another characteristic feature of these elements is the reproducible biochemical trademark exemplified as protein binding or specific chromatin structure. Using 1640 data sets from 147 different cell types complemented with all ENCODE data regarding candidate regions from genome-wide association studies (GWAS) and evolutionarily constrained territories revealed important components about the function and organization of the human genome. The main conclusions drawn by the ENCODE consortium were the following: “1, More than 80% of the human genome is associated with at least one biochemical signature in one particular cell type. 2, Classification of the genome into functionally different chromatin states implies an initial set of approximately 400,000 enhancer-like regions and more than 70,000 promoter-like elements. 3, Quantitative correlation of RNA production versus chromatin marks and transcription factor binding on the promoter regions indicate that RNA expression is mostly dependent on the functionality of the promoter. 4, ENCODE annotated at least as much functional non-coding DNA sequences as the protein-coding genes. 5, Single nucleotide polymorphisms (SNPs) associated with disease phenotypes determined by GWAS enriched in non-coding functional elements annotated by the consortium” [60]. Based on these considerations it is obvious that the non-coding part of the genome is full of functional and disease associated regulatory elements. The basis to connect these to the distal target genes remained unexplored. Gene promoters and their *cis*-acting elements can participate in looping that is involved in gene regulation [17]. In order to link genes and their putative regulatory regions, chromosome conformation capture carbon copy (5C) is carried out [13]. 5C interaction maps are generated from three cell lines and these results were integrated with the ENCODE data. Merging these datasets revealed more than 1000 interactions between enhancers, CCCTC-binding factor (CTCF) bound sites and promoters in each cell lines. Significant correlations are observed between gene expression and the existence of promoter-enhancer interaction and the presence of transcripts originating from enhancers (enhancer RNAs, eRNAs). Interestingly, unlike the reported functions of CTCF as an insulator, long-range interactions are not blocked by CTCF-cohesin co-bound sites demonstrating that many of these sites are not demarcated physically insulated gene domains. The fact that only 7% of the loops are detected with the closest gene, suggests that genomic proximity is not necessarily a good predictor for long-range interactions [61]. This study clearly demonstrates that cis-acting elements communicate with their targeted promoters via looping, nevertheless if one would like to reveal the entire interaction map between these elements, one needs a more robust method capable of detecting all co-operations. Although 5C is a very powerful method to map the interactions, it is limited to a single locus. The usage of ChIA-PET solved this issue. In order to link the regulatory elements to their targets, DNase I hypersensitive sites (DHS) are resolved leading to the determination of open chromatin landscape of several cell lines representing the human genome. These regions were then aligned with RNAPII ChIA-PET results gathering all the participants of RNAPII dependent open chromatin interactions. This large-scale interaction analysis confirmed that cooperation between DHS sites and promoters are markedly enriched. Surprisingly, this kind of integration revealed that approximately half of the DHS sites are detectable in the close proximity of more than one promoter. These results suggest that the human cis-acting element network is more complex than anticipated [62].

Taken together, ENCODE identified an enormous amount of functional elements in the human genome and provides a beneficial resource for the field. On the other hand the data presented is greatly enlarged our understanding about the functionality of the human genome directing us toward new challenges regarding *cis*-acting element annotation and how these act genome-wide. The functional significance of such detected interactions is not known at all.

Pioneering, bookmarking and higher order chromatin structure

In the previous sections we summarized the knowledge regarding the features of cis-acting elements and how one can recognize them in the very complex mammalian genome. Each of the genes encoded in the human or mammalian genetic material is likely to have many distinct *cis*-acting elements spread across tens to hundreds of kilobases. These regulatory sequences act in concert to fine-tune gene expression in a highly tissue and signal specific manner. What kind of factors dictate the functionality of such elements in the genome leading to differential gene expression patterns in various cell types? Next we focus on the pioneering and bookmarking factor concept and on the mechanistic determinants of higher order chromatin structure contributing to cell specific transcription. From a functional point of view there are at least three categories of enhancers: 1. Potential enhancers not binding its cognate transcription factor. 2. Non-active enhancers, binding a particular transcrtiption factor, but not participating in enhancement of transcription and 3. Active enhancers, binding the required transcription factor and activating transcription. Molecularly, the diversity is likely to be greater. These need to be sorted out, linked to regulated gene(s) and functionally validated.

From a mechanistic point of view enhancer elements must be prepared before activation. It has been shown that the forkhead transcription factor A (FoxA) binds to inactive genomic regions and by remodeling the nucleosome pattern, it is able to recruit other transcription factors leading to enhancer activation. The first two pioneer factors (FoxA, GATA) have been described in the liver differentiation program. By definition pioneer factors hold the property to bind nucleosomal DNA and compact chromatin, and remain bound during mitosis. FoxA is a typical pioneer factor, which is capable of transform the actual enhancer to a state called „poised" for activation [63]. This state of the enhancers renders them for rapid activation once the specific signal or its downstream effector appears. Other transcription factors have been also described with pioneer properties as reviewed in [64]. According to pioneer factors, the prevalent view is that during differentiation these factors shape the active *cis*-regulatory element network, thus contributes to the acquisition of cell identity. More precisely, this concept describes that the regulatory landscape evolved through the differentiation program determines the sites where transcriptional regulation occurs. Interestingly, in 2013 Ostuni et al. reported the existence of the so-called latent enhancers. These *cis*-elements are not bound by the lineage-specific transcription factor PU.1 in terminally differentiated macrophages and do not show the characteristic histone profiles of enhancers. Importantly, upon stimulation by an activating signal (IL4, INFg, TGFb), their histone profile can suddenly change due to the binding of the downstream effectors and leads to the subsequent binding of PU.1. After stimulation, in the absence of the activating signal most of them do not return to the latent state, instead remained marked by H3K4me1 and upon exposure they mediate a faster and stronger response, thus provides an epigenetic memory to the cell [65]. It remains to be seen how wide spread is this mechanism among different cell types and cell types.

Taken together, the picture seems quite complex. Cell type specific pioneer factors exist and at least in part shape the *cis*-acting element landscape across cell types. Other factors affecting the architecture of the genome have been also shown to be important in establishing the functional regulatory elements as reviewed in [66]. At the same time it is also known that not all lineage specific transcription factors have chromatin remodeling activities associated with. Therefore these are more appropriate to be called as bookmarking factors. It remains to be discovered how these factors i.e. PU.1 establish cellular memory mechanistically.

It has been shown that each chromosome has its own physical location in the nucleus as reviewed [55]. Chromosomal territories are functionally different and spatially separated, but what are the determinants of this higher-order chromatin structure, which is implicated in the regulation of gene expression and also responsible for cell autonomous transcriptomes?

This line of investigations started when insulators were first recognized in vertebrates [67]. Insulators were first described because of their ability to block enhancer function. Later on it has been shown that CTCF allows these insulator sequences to function as separating enhancer/promoter interactions and also active/passive gene domains [68]. CTCF is a transcription factor possessing elven zinc-fingers and is ubiquitously expressed in higher eukaryotes. The function of CTCF was further clarified showing that it functions with the multiprotein cohesin complex containing the following subunits: SMC1, SMC3, RAD21, SA1, SA2 [69]. Cohesin has a ring-like shape with a diameter of approximately 40 nm. This important feature of cohesin makes it suitable to handle the chromatin fiber as shown in sister chromatid cohesion [70]. According to the described features of these proteins it is conceivable that they are crucial components of shaping the higher-order chromatin structure.

CTCF had long been thought to contribute to the structural organization of the genome, but its long-range interaction mediating effect has remained elusive until it has been linked to cohesin on the mouse Infg locus [71]. This study was the very first to show that both CTCF and cohesin are indispensable for genomic interactions. Others also reported that cohesin depletion leads to diminished promoter-enhancer interactions in embryonic stem cells [72] and in thymocites [73]. A series of genome-wide studies show that CTCF and cohesin co-occupy regions in the genome [74], [75]. The extensive interaction between these factors may explain how CTCF separates functionally different domains.

Recently, contact mapping of chromosomes determined by NGS based methods revealed the topological domain structure of the genome [16], [13]. These domains contain multiple genes and possess differential gene expression activity and epigenetic pattern. Presumably, these domains serve as fundamental building blocks that support active and passive chromosomal architectures. It has been shown that the anchoring points of chromatin loops, organizing the domain structure are enriched for CTCF and cohesin binding sites [76]. Other studies have reported that CTCF-cohesin co-bound regions mediate the looping events surrounding promoter enhancer elements, while those regions occupied only by cohesin are responsible for enhancer-promoter interactions [72]. Based on these results, several studies showed that CTCF and cohesin are required to maintain topological domain structures, interchromosomal interactions and enhancer-promoter interactions [77], [78], [79] (Figure 1C, 2B). Perturbation of the cohesin complex has been shown to affect gene expression involving not only the cohesin bound genes, but also those that are free of cohesin suggesting its function in preserving topological domain structure [78]. According to these results CTCF and cohesion is likely to play key roles in proper gene regulation, although their genomic binding sites suggest that their effects on gene expression are not cell type specific because the CTCF cistrome is largely invariant between cell types. As part of the ENCODE project a study compared CTCF binding sites from 19 different human primary and immortal cell lines. Surprisingly, the result shows that there is plasticity in CTCF binding across cell types indicative of strong cell-selective regulation of CTCF binding. Using massively parallel bisulfite sequencing the authors showed that approximately 40% of variable CTCF binding is due to differential methylation states at two specific points of the binding motif. Strikingly they could demonstrate that CTCF binding is dramatically different between primary and immortal cell lines. The latter harbors widespread disruption of CTCF sites associated with increased methylation [77].

Taken together, CTCF-cohesin co-bound sites appear to be responsible, at least in part, for the configuration of topological domain structure. By shaping the genome architecture these factors significantly contribute to the regulation of gene expression also in a cell type specific manner.

Is there a function for enhancer RNAs?

Recent advances in genomic technologies made the surprising finding that active enhancers are transcribed into RNA molecules, called enhancer RNAs (eRNA). The first results describing the existence of eRNAs originated from the locus control region (LCR) of the beta-globin gene clusters [80], [81]. The fact that there is pervasive transcription on enhancer elements came with the advent of total RNA sequencing, showing that in neuronal activity regulated and T-cell specific enhancers are transcribed [48], [49]. More studies were published recently in the field using GRO-Seq in various cells and species, which have clearly demonstrated that enhancers are transcribed to eRNAs and to a given stimuli, the activation dynamics of these elements are similar to their targeted genes [50], [82].

There is a debate on the field about eRNA function. The question is very simple: Are these functionally relevant in gene expression regulation or eRNAs are just merely byproducts of gene transcription? Recently, several studies published using novel methods to test the functionality of enhancer-derived transcripts. Specific degradation of eRNAs using either RNA interference or antisense oligonucleotides demonstrated that the expression of the adjacent gene targeted by the enhancer is reduced [83], [84], [85], [86]. Two out of the four studies also performed an eRNA tethering assay in a reporter system. Interestingly, eRNA connected to either the promoter [85] or the enhancer [83] was capable of increasing the expression of the reporter gene. As further evidence, Lam et al. showed that by integrating various sizes of genomic regions from an enhancer to a reporter vector could differentially affect the activity of the reporter gene. If the core sequence was cloned containing the TFBS only, they detected increased reporter activity compared to a reporter plasmid carrying random DNA sequence. Surprisingly, the reporter encompassing the core and the entire eRNA-coding sequence had the highest transcriptional output. Inverting the eRNA-coding region supported the most striking result because changing the sequence of the eRNA diminished its boosting effect, suggesting that the sequence of these short RNAs can be very important in the context of gene regulation [84]. According to these results, it seems that eRNAs possess regulatory function, but the question is how. What is the molecular mechanism through which these small molecules reach the heart of transcription?

It has been demonstrated that enhancers actively participating in looping with their targeted promoters acquire higher level of eRNAs [87], [61]. Based on these studies the predicted model must be that eRNAs somehow secure the contact between enhancers and their corresponding promoters. Nuclear receptors like estrogen receptor alpha (ERa) have been shown to bind enhancer elements and in the presence of the agonist in this particular case oestrogen, these *cis*-acting elements are anchored at target gene promoters through long-range chromatin interactions determined by ChIA-PET [17]. Recently, a very important finding came to light showing that knockdown of eRNAs immediately next to ERa bound enhancers reduced enhancer-promoter interactions and resulted in a reduced expression level of the corresponding gene. Potentially these ERa-mediated eRNAs are taking part in the modulation of looping, which was further supported by the fact that eRNAs could pull-down the subunits of the cohesin complex. On the other hand RNA immunoprecipitation (RIP) assays performed against RAD21 showing that eRNAs could enrich in the cohesin complex [83] which has been shown to control enhancer-promoter interactions [72]. Furthermore, targeted degradation of eRNAs by RNA interference led to the inhibition of oestrogen dependent RAD21 recruitment at several ERa bound enhancers. Strikingly, knockdown of RAD21 almost fully diminished the interaction on the gene loci NRIP1 and GREB1 between the enhancers and their corresponding promoters. In addition, knockdown of SMC1, another component of the cohesin complex, almost completely abolished the oestrogen mediated gene activation program [83]. Thus eRNAs may participate in the process of looping by initiating or stabilizing the interactions of enhancer-promoter pairs.

In an independent study by Hah et al. showed that inhibiting RNAPII elongation (by flavopiridol) has no significant effect on the looping examined on the P2RY2 and GREB1 loci using 3C [88]. These results put a little bit of contradiction into the field, but the difference may be explained by the usage of different experimental systems (eRNA silencing vs. Pharmacological inhibition of RNAPII elongation) or these may reflect different mechanisms at different gene loci [43].

Overall these studies suggest that at least in some cases eRNAs possess regulatory function and contributes to gene expression regulation. Clearly, further studies are needed to clarify their roles.

Enhancers and disease

Previously we summarized the main features of an enhancer element and also delineate their annotation to their corresponding genes, than we dealt with eRNA function. In this part we sum up those studies, which connect single nucleotide polymorphisms (SNPs) with cis-acting elements supporting the notion that SNPs in the regulatory regions may cause disease phenotypes and thus contributes to evolution. Disruptions in chromosomal regions not harboring coding genes provided the first evidence that mutations in the non-coding part of the genome may contribute to disease development. 1.8% of the identified point mutations fell in the non-coding part of the genome based on the Human Gene Mutation Database.

Mutations in the cis-acting elements can be easily annotated to a given disease phenotype if they fulfill one or more of the following: 1, Genetic evidence is present to link the phenotype to a known disease locus. 2, Validated chromosomal anomaly (deletion, amplification, rearrangement) can be annotated to a known disease gene. 3, Resulting phenotype is very similar to the phenotypic change caused by a mutation in the coding region of a known disease gene. 4, Disease associated variation accounts for all or a significant fraction of disease risk [89].

One of the earliest studies mentioned the regulatory element mutations of the PAX6 locus involved in Aniridia. Aniridia is characterized by the absence of iris and is mostly due to mutations occurred in the coding sequence of the PAX6 gene [90]. However a fraction of the cases are not caused by mutations in the coding sequence. It has been shown that the downstream genomic region is full of rearrangements. The most distal point harboring the mutation lies approximately 125kb from the last exon of PAX6 and fall in the intronic region of the ubiquitously expressed gene ELP4, although haploinsufficiency for ELP4 has been shown not to contribute to the disease phenotype. YAC-based transgene experiments in mice revealed an approximately 80kb long genomic region containing series of DHSs. Later on, it has been described that these *cis*-acting elements are required for PAX6 expression [91].

Another striking example has been demonstrated with the POU3F4 gene. Mutations occurred in the coding region responsible for X-linked deafness type 3, however a smaller group of cases was identified that lack the gene variation. Interestingly, approximately 900kb upstream from the gene’s TSS, a very important 2kb element has been described to overlap with an otic vesicle enhancer, likely regulating POU3F4 [92].

By these experiments it has been shown that regulatory elements function over long distances, they can reside in other transcriptional units and their mutations can cause disease phenotypes. The listed diseases are typically inherited in a Mendelian manner, although mutations present in the non-coding regions of the genome where they associated with non-Mendelian diseases.

The rapid evolution of genotyping technologies has resulted in GWAS data that generally imply a powerful role for regulatory variation in common genetic disorders [93]. Recently, meta-analysis of approximately 1200 SNPs representing the most significant association with disease phenotypes has been done. Surprisingly 40% of these falls into the non-coding part of the genome suggesting that disease causing mutations may act on enhancer elements [94].

Several studies successfully demonstrated roles for non-coding mutations in disease risk. One of the earliest was identified in the intronic region of the RET gene contributing to Hirschprung disease [95]. Recently, it has been shown that a mutation reside in the non-coding region belong to the IRF6 gene is a risk factor of cleft lip associated with Van der Woude Syndrome [96]. Similar observations have been made in several other cases reviewed in [89].

Overall, the vast amount of data unequivocally support the notion that cis-acting elements are crucial components of proper gene regulation and their mutations give rise to various pathological conditions. Importantly, until now no one could investigate the effect of these genetic variations on the functional cis-acting element toolkit in a special cell type.

Recently, Heinz et al. used different mouse strains looking for differential binding of transcription factors caused by natural genetic variation. In this study they could delineate the strain specific differences in functional enhancer usage in macrophages. Most of the cells express hundreds of transcription factors to control the non-coding part of the genome, thus shaping the cell specific transcriptome. In macrophages, transcription factors responsible for marking regulatory elements are PU.1, C/EBPs and the AP-1 family members. They showed that in macrophages these lineage determining transcription factors (LDTF) collaboratively occupy the 70% of the active enhancer elements. Surprisingly, if an SNP can be detected in the PU.1 motif leading to diminished PU.1 binding, it negatively affected the binding of the other two LDTFs. Conversely, if the binding motif of C/EBP or AP-1 harbors the mutation, PU.1/AP-1 and PU.1/C/EBP binding also diminished. In addition, H3K4me2 and H3K27ac markers of active enhancers are also abolished, meaning that these enhancers were not functional any longer. These results provided a definitive answer to the question: How enhancer function and transcription factor binding is lost where there are no mutations in its binding motif? Based on these the answer is the requirement for collaborative binding. Interestingly, further examination of the strains in the context of a signal specific transcription factor, NF-κB led to the observation that mutations occurred in the LDTF motifs are approximately three times more likely to result in decreased NF-κB binding, than mutations occurred in the NF-κB motif, showing the importance of LDTFs in chromatin priming/remodeling as previously mentioned [97].

The authors claim that this collaborative binding model can be very useful if one determines the LDTFs in a given cell type and merge these binding sites with the annotated genetic variations to pinpoint the potential disease-causing variants. The challenge will be in the future to expand these studies and use them in different model systems to understand disease-causing natural genetic variations [97].

Conclusions and future directions

Transcriptional control of cell specific gene expression is carried out through complex molecular mechanisms. Recent works harnessing the advantage of the integration of high-throughput sequencing data led to the identification of active enhancer elements in various cell types. Enhancers seems to be highly tissue specific, explained by the presence of different pioneering factors establishing the cis-acting element repertoire and the higher order chromatin structure shaped partly by CTCF and CTCF/cohesin to our best knowledge. Genomic regions are partitioned into functionally distinct gene domains in which regulatory elements may act on various genes. The main molecular mechanism of gene expression regulation often carried out over long distances covered by chromatin loops between the regulator and the regulated gene.

Functional enhancers actively engaged in looping and possess the ability to be transcribed to eRNAs. These short RNA pieces originated from enhancers seems to be important components of those protein complexes taking part in looping. Importantly, this process at least in general terms is indispensable for efficient assembly of transcription initiation machinery on target gene promoters. Although these studies greatly improved our view about gene regulation, still inherently descriptive harboring the disadvantage that we could not render a clear biological function to these cis-elements unless a clear human disease phenotype is connected to them by knowing the enhancer SNP location. In the future probably most of the enhancers will be cut out by using genome-editing methods, for instance TALEN [98] or CRISPR [99]. Genome editing is expected to fulfill the expectation that the function of each element will be identified and by these means create the opportunity to selectively target each of them to cure diseases. The results obtained from the silencing of eRNAs are also promising, showing that targeting an enhancer transcript derived from an active cis-acting element may has an effect on the target gene expression, thus serves as a remote control for genes. The problems with these are: 1. Their specificity is questionable because one enhancer can be responsible for the regulation of several genes. 2, Presence of shadow enhancers [100] can compensate the effect of the original enhancer. 3, Another challenge can be to hit the target specifically under physiological setting.

All together integration based genomics holds the potential to reveal many-many new drug targets and to identify most of the disease causing mutations, which can be located in the non-coding part of the genome and so far was very cumbersome to seek out.

Figure legends:

Figure 1. Transcription regulation in prokaryotes and eukaryotes and the complexity of gene regulation in eukaryotic cells.

A, Prokaryotic gene expression is based on operons in which a gene cluster is under the control of two genomic (promoter) regions. The first promoter located at the 5' end is responsible for the expression of the regulator protein, which in turn silences the whole operon via binding to the operator region. In the presence of an activating stimuli the regulator cannot bind to the operator region, thus the second promoter will be active and leads to the efficient expression of the enzyme coding genes producing polycistronic RNA molecules encoding more protein products.

B, Eukaryotic genes are typically regulated by cis-acting elements located in the non-coding part of the genome. These elemets can be located far away from their target genes, even in an other chromosome, thus it is challenging to pair them with their genes. RNA synthesized from a eukaryotic gene is monocistromic and undergoes the process called splicing, in which the intronic regions are excluded from the nascent transcript before translation.

C, Eukaryotic transcription might be coordinately regulated in the so called transcription factories. These subnuclear compartments might be, in part, stabilized by the CTCF/Cohesin protein complexes and permit the expression of genes in a well-coordinated manner, mechanically connecting genes regulated by the same signal, but residing on distinct chromosomes. A hypothetical scenario is depicted in which several genomic regions on distinct chromosomes are linked by CTCF/Cohesin interactions and co-localize in the nucleus forming a transcription factory. Interchromosomal interactions are marked by asterisks.

Figure 2. Schematic representation of enhancer function and the establisment of cell type specific gene expression program.

A, Enhancer-gene pairs are located within a topological domain bordered by CTCF/Cohesin protein complexes. Enhancers may be located long distances away relative to the TSS of the regulated gene. Upon signal-specific enhancer activation transcription factor binding and chromatin remodeling occurs leading to the deposition of enhancer marks for instance H3K27ac by the coactivator complex. Enhancers activated by the stimuli then get in to the close proximity of the promoter region potentially by an eRNA-dependent mechanism which also relies on the cohesin complex. After these steps the enhancers can initiate or boost the expression of the target by supporting the assembly of the general transcription factors and the mediator complex along with the deposition of the active TSS mark H3K4me3.

B, Cell type specific gene expression program is established by the existence of various cell type specific transcription factors (pioneering/bookmarking factors) which can shape the functional cis-acting element landscape of the genome. Though the higher order chromatin structure is also implicated in the designation of active gene domains. In cell A, the functional gene domain is demarked by CTCF/cohesin cobound regions and the active enhancer element is occupied by the pioneering/bookmarking or the lineage determining factor, thus support an entry point to the signal dependent transcription factor. In cell B, the same gene domain is active, but because of the presence of a different pioneering/bookmarking factor it uses a different regulatory element, which in turn lead to a differential gene expression output, while in cell C the chromatin conformation of the locus is reshaped due to the differential binding of CTCF/Cohesin, hence a new enhancer unit can get in to the proximity of Gene A possessing another cell type specific pioneering/bookmarking factor with the ability to push the element into a poised for activation state.

Acknowledgement

The authors would like to acknowledge the members of the Nagy laboratory and Dr. Balint L. Balint for discussions and comments on the manuscript. L.N is supported by a grant from the Hungarian Scientific Research Fund (OTKA K100196) and TÁMOP-4.2.2.A-11/1/KONV-2012-0023 VÉD-ELEM and G.N. is a recipient of TÁMOP 4.2.4. A/2-11-1-2012-0001/A2-JÁDJ-13

both implemented through the New Hungary Development Plan co-financed by the European Social Fund and the European Regional Development Fund.

References

1. Gregory, T. R. (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma, *Biological reviews of the Cambridge Philosophical Society.* 76, 65-101.
2. Stanford, W. L., Cohn, J. B. & Cordes, S. P. (2001) Gene-trap mutagenesis: past, present and beyond, *Nature reviews Genetics.* 2, 756-68.
3. Banerji, J., Rusconi, S. & Schaffner, W. (1981) Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences, *Cell.* 27, 299-308.
4. Augenlicht, L. H. & Kobrin, D. (1982) Cloning and screening of sequences expressed in a mouse colon tumor, *Cancer research.* 42, 1088-93.
5. Kulesh, D. A., Clive, D. R., Zarlenga, D. S. & Greene, J. J. (1987) Identification of interferon-modulated proliferation-related cDNA sequences, *Proceedings of the National Academy of Sciences of the United States of America.* 84, 8453-7.
6. Augenlicht, L. H., Taylor, J., Anderson, L. & Lipkin, M. (1991) Patterns of gene expression that characterize the colonic mucosa in patients at genetic risk for colonic cancer, *Proceedings of the National Academy of Sciences of the United States of America.* 88, 3286-9.
7. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science.* 270, 467-70.
8. Park, P. J. (2009) ChIP-seq: advantages and challenges of a maturing technology, *Nature reviews Genetics.* 10, 669-80.
9. Kapranov, P., Willingham, A. T. & Gingeras, T. R. (2007) Genome-wide transcription and the implications for genomic organization, *Nature reviews Genetics.* 8, 413-23.
10. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. (2002) Capturing chromosome conformation, *Science.* 295, 1306-11.
11. Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. & de Laat, W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C), *Nature genetics.* 38, 1348-54.
12. Stadhouders, R., Kolovos, P., Brouwer, R., Zuin, J., van den Heuvel, A., Kockx, C., Palstra, R. J., Wendt, K. S., Grosveld, F., van Ijcken, W. & Soler, E. (2013) Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions, *Nature protocols.* 8, 509-24.
13. Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D. & Dekker, J. (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements, *Genome research.* 16, 1299-309.
14. van de Werken, H. J., Landan, G., Holwerda, S. J., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B. A., Verstegen, M. J., de Wit, E., Tanay, A. & de Laat, W. (2012) Robust 4C-seq data analysis to screen for regulatory DNA interactions, *Nature methods.* 9, 969-72.

15. Fraser, J., Ethier, S. D., Miura, H. & Dostie, J. (2012) A Torrent of data: mapping chromatin organization using 5C and high-throughput sequencing, *Methods in enzymology.* 513, 113-41.
16. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S. & Dekker, J. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome, *Science.* 326, 289-93.
17. Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., Chew, E. G., Huang, P. Y., Welboren, W. J., Han, Y., Ooi, H. S., Ariyaratne, P. N., Vega, V. B., Luo, Y., Tan, P. Y., Choy, P. Y., Wansa, K. D., Zhao, B., Lim, K. S., Leow, S. C., Yow, J. S., Joseph, R., Li, H., Desai, K. V., Thomsen, J. S., Lee, Y. K., Karuturi, R. K., Herve, T., Bourque, G., Stunnenberg, H. G., Ruan, X., Cacheux-Rataboul, V., Sung, W. K., Liu, E. T., Wei, C. L., Cheung, E. & Ruan, Y. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome, *Nature.* 462, 58-64.
18. Horike, S., Cai, S., Miyano, M., Cheng, J. F. & Kohwi-Shigematsu, T. (2005) Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome, *Nature genetics.* 37, 31-40.
19. Trapnell, C., Pachter, L. & Salzberg, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics.* 25, 1105-11.
20. Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C. & Zhang, J. (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data, *PLoS computational biology.* 9, e1003326.
21. Bailey, T. L., Baker, M. E. & Elkan, C. P. (1997) An artificial intelligence approach to motif discovery in protein sequences: application to steriod dehydrogenases, *The Journal of steroid biochemistry and molecular biology.* 62, 29-44.
22. Knuppel, R., Dietze, P., Lehnberg, W., Frech, K. & Wingender, E. (1994) TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins, *Journal of computational biology : a journal of computational molecular cell biology.* 1, 191-8.
23. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. & Lenhard, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles, *Nucleic acids research.* 32, D91-4.
24. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities, *Molecular cell.* 38, 576-89.
25. Rhee, H. S. & Pugh, B. F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution, *Cell.* 147, 1408-19.
26. Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K. & Peng, W. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data, *Bioinformatics.* 25, 1952-8.
27. Yuan, G. C., Liu, Y. J., Dion, M. F., Slack, M. D., Wu, L. F., Altschuler, S. J. & Rando, O. J. (2005) Genome-scale identification of nucleosome positions in S. cerevisiae, *Science.* 309, 626-30.
28. Lee, W., Tillo, D., Bray, N., Morse, R. H., Davis, R. W., Hughes, T. R. & Nislow, C. (2007) A high-resolution atlas of nucleosome occupancy in yeast, *Nature genetics.* 39, 1235-44.

29. Sun, W., Xie, W., Xu, F., Grunstein, M. & Li, K. C. (2009) Dissecting nucleosome free regions by a segmental semi-Markov model, *PloS one.* 4, e4721.
30. Nagy, G., Daniel, B., Jonas, D., Nagy, L. & Barta, E. (2013) A novel method to predict regulatory regions based on histone mark landscapes in macrophages, *Immunobiology.* 218, 1416-27.
31. Gaulton, K. J., Nammo, T., Pasquali, L., Simon, J. M., Giresi, P. G., Fogarty, M. P., Panhuis, T. M., Mieczkowski, P., Secchi, A., Bosco, D., Berney, T., Montanya, E., Mohlke, K. L., Lieb, J. D. & Ferrer, J. (2010) A map of open chromatin in human pancreatic islets, *Nature genetics.* 42, 255-9.
32. Auerbach, R. K., Euskirchen, G., Rozowsky, J., Lamarre-Vincent, N., Moqtaderi, Z., Lefrancois, P., Struhl, K., Gerstein, M. & Snyder, M. (2009) Mapping accessible chromatin regions using Sono-Seq, *Proceedings of the National Academy of Sciences of the United States of America.* 106, 14926-31.
33. Weber, J., Jelinek, W. & Darnell, J. E., Jr. (1977) The definition of a large viral transcription unit late in Ad2 infection of HeLa cells: mapping of nascent RNA molecules labeled in isolated nuclei, *Cell.* 10, 611-6.
34. Core, L. J., Waterfall, J. J. & Lis, J. T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters, *Science.* 322, 1845-8.
35. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. (2003) Scanning human gene deserts for long-range enhancers, *Science.* 302, 413.
36. Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E. & de Graaff, E. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly, *Human molecular genetics.* 12, 1725-35.
37. Malik, S. & Roeder, R. G. (2010) The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation, *Nature reviews Genetics.* 11, 761-72.
38. Clapier, C. R. & Cairns, B. R. (2009) The biology of chromatin remodeling complexes, *Annual review of biochemistry.* 78, 273-304.
39. Mito, Y., Henikoff, J. G. & Henikoff, S. (2007) Histone replacement marks the boundaries of cis-regulatory domains, *Science.* 315, 1408-11.
40. Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E. & Ren, B. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome, *Nature genetics.* 39, 311-8.
41. Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K. & Felsenfeld, G. (2009) H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions, *Nature genetics.* 41, 941-5.
42. Bedford, D. C., Kasper, L. H., Fukuyama, T. & Brindle, P. K. (2010) Target gene context influences the transcriptional requirement for the KAT3 family of CBP and p300 histone acetyltransferases, *Epigenetics : official journal of the DNA Methylation Society.* 5, 9-15.
43. Lam, M. T., Li, W., Rosenfeld, M. G. & Glass, C. K. (2014) Enhancer RNAs and regulated transcriptional programs, *Trends in biochemical sciences.* 39, 170-182.
44. Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D. & Ren, B. (2005) A high-resolution map of active promoters in the human genome, *Nature.* 436, 876-80.
45. Kouzarides, T. (2007) Chromatin modifications and their function, *Cell.* 128, 693-705.

46. Lupien, M., Eeckhoute, J., Meyer, C. A., Wang, Q., Zhang, Y., Li, W., Carroll, J. S., Liu, X. S. & Brown, M. (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription, *Cell.* 132, 958-70.
47. Lin, Y. C., Jhunjhunwala, S., Benner, C., Heinz, S., Welinder, E., Mansson, R., Sigvardsson, M., Hagman, J., Espinoza, C. A., Dutkowski, J., Ideker, T., Glass, C. K. & Murre, C. (2010) A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate, *Nature immunology.* 11, 635-43.
48. Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G. & Greenberg, M. E. (2010) Widespread transcription at neuronal activity-regulated enhancers, *Nature.* 465, 182-7.
49. Koch, F., Fenouil, R., Gut, M., Cauchy, P., Albert, T. K., Zacarias-Cabeza, J., Spicuglia, S., de la Chapelle, A. L., Heidemann, M., Hintermair, C., Eick, D., Gut, I., Ferrier, P. & Andrau, J. C. (2011) Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters, *Nature structural & molecular biology.* 18, 956-63.
50. Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M. U., Ohgi, K. A., Glass, C. K., Rosenfeld, M. G. & Fu, X. D. (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA, *Nature.* 474, 390-4.
51. Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B. L., Couronne, O., Eisen, M. B., Visel, A. & Rubin, E. M. (2006) In vivo enhancer analysis of human conserved non-coding sequences, *Nature.* 444, 499-502.
52. Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M. & Pennacchio, L. A. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers, *Nature.* 457, 854-8.
53. Carroll, J. S., Liu, X. S., Brodsky, A. S., Li, W., Meyer, C. A., Szary, A. J., Eeckhoute, J., Shao, W., Hestermann, E. V., Geistlinger, T. R., Fox, E. A., Silver, P. A. & Brown, M. (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1, *Cell.* 122, 33-43.
54. Barish, G. D., Yu, R. T., Karunasiri, M. S., Becerra, D., Kim, J., Tseng, T. W., Tai, L. J., Leblanc, M., Diehl, C., Cerchietti, L., Miller, Y. I., Witztum, J. L., Melnick, A. M., Dent, A. L., Tangirala, R. K. & Evans, R. M. (2012) The Bcl6-SMRT/NCoR cistrome represses inflammation to attenuate atherosclerosis, *Cell metabolism.* 15, 554-62.
55. de Wit, E. & de Laat, W. (2012) A decade of 3C technologies: insights into nuclear organization, *Genes & development.* 26, 11-24.
56. 1. West, A. G. & Fraser, P. (2005) Remote control of gene transcription, *Human molecular genetics.* 14 Spec No 1, R101-11.
57. Lander, E. S.Linton, L. M.Birren, B.Nusbaum, C.Zody, M. C.Baldwin, J.Devon, K.Dewar, K.Doyle, M.FitzHugh, W.Funke, R.Gage, D.Harris, K.Heaford, A.Howland, J.Kann, L.Lehoczky, J.LeVine, R.McEwan, P.McKernan, K.Meldrim, J.Mesirov, J. P.Miranda, C.Morris, W.Naylor, J.Raymond, C.Rosetti, M.Santos, R.Sheridan, A.Sougnez, C.Stange-Thomann, N.Stojanovic, N.Subramanian, A.Wyman, D.Rogers, J.Sulston, J.Ainscough, R.Beck, S.Bentley, D.Burton, J.Clee, C.Carter, N.Coulson, A.Deadman, R.Deloukas, P.Dunham, A.Dunham, I.Durbin, R.French, L.Grafham, D.Gregory, S.Hubbard, T.Humphray, S.Hunt, A.Jones, M.Lloyd, C.McMurray, A.Matthews, L.Mercer, S.Milne, S.Mullikin, J. C.Mungall, A.Plumb, R.Ross,

M.Shownkeen, R.Sims, S.Waterston, R. H.Wilson, R. K.Hillier, L. W.McPherson, J. D.Marra, M. A.Mardis, E. R.Fulton, L. A.Chinwalla, A. T.Pepin, K. H.Gish, W. R.Chissoe, S. L.Wendl, M. C.Delehaunty, K. D.Miner, T. L.Delehaunty, A.Kramer, J. B.Cook, L. L.Fulton, R. S.Johnson, D. L.Minx, P. J.Clifton, S. W.Hawkins, T.Branscomb, E.Predki, P.Richardson, P.Wenning, S.Slezak, T.Doggett, N.Cheng, J. F.Olsen, A.Lucas, S.Elkin, C.Uberbacher, E.Frazier, M., et al. (2001) Initial sequencing and analysis of the human genome, *Nature.* 409, 860-921.
58. Consortium, E. P. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project, *Science.* 306, 636-40.
59. Maher, B. (2012) ENCODE: The human encyclopaedia, *Nature.* 489, 46-8.
60. Consortium, E. P., Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C. & Snyder, M. (2012) An integrated encyclopedia of DNA elements in the human genome, *Nature.* 489, 57-74.
61. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. (2012) The long-range interaction landscape of gene promoters, *Nature.* 489, 109-13.
62. Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B. K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E. & Stamatoyannopoulos, J. A. (2012) The accessible chromatin landscape of the human genome, *Nature.* 489, 75-82.
63. Watts, J. A., Zhang, C., Klein-Szanto, A. J., Kormish, J. D., Fu, J., Zhang, M. Q. & Zaret, K. S. (2011) Study of FoxA pioneer factor at silent genes reveals Rfx-repressed enhancer at Cdx2 and a potential indicator of esophageal adenocarcinoma development, *PLoS genetics.* 7, e1002277.
64. Magnani, L., Eeckhoute, J. & Lupien, M. (2011) Pioneer factors: directing transcriptional regulators within the chromatin environment, *Trends in genetics : TIG.* 27, 465-74.
65. Ostuni, R., Piccolo, V., Barozzi, I., Polletti, S., Termanini, A., Bonifacio, S., Curina, A., Prosperini, E., Ghisletti, S. & Natoli, G. (2013) Latent enhancers activated by stimulation in differentiated cells, *Cell.* 152, 157-71.
66. Ong, C. T. & Corces, V. G. (2014) CTCF: an architectural protein bridging genome topology and function, *Nature reviews Genetics.* 15, 234-46.
67. Chung, J. H., Whiteley, M. & Felsenfeld, G. (1993) A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila, *Cell.* 74, 505-14.
68. Bell, A. C., West, A. G. & Felsenfeld, G. (1999) The protein CTCF is required for the enhancer blocking activity of vertebrate insulators, *Cell.* 98, 387-96.
69. Xiao, T., Wallace, J. & Felsenfeld, G. (2011) Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity, *Molecular and cellular biology.* 31, 2174-83.
70. Nasmyth, K. & Haering, C. H. (2009) Cohesin: its roles and mechanisms, *Annual review of genetics.* 43, 525-58.

71. Hadjur, S., Williams, L. M., Ryan, N. K., Cobb, B. S., Sexton, T., Fraser, P., Fisher, A. G. & Merkenschlager, M. (2009) Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus, *Nature.* 460, 410-3.
72. Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., Taatjes, D. J., Dekker, J. & Young, R. A. (2010) Mediator and cohesin connect gene expression and chromatin architecture, *Nature.* 467, 430-5.
73. Seitan, V. C., Hao, B., Tachibana-Konwalski, K., Lavagnolli, T., Mira-Bontenbal, H., Brown, K. E., Teng, G., Carroll, T., Terry, A., Horan, K., Marks, H., Adams, D. J., Schatz, D. G., Aragon, L., Fisher, A. G., Krangel, M. S., Nasmyth, K. & Merkenschlager, M. (2011) A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation, *Nature.* 476, 467-71.
74. Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., Cobb, B. S., Yokomori, K., Dillon, N., Aragon, L., Fisher, A. G. & Merkenschlager, M. (2008) Cohesins functionally associate with CTCF on mammalian chromosome arms, *Cell.* 132, 422-33.
75. Rubio, E. D., Reiss, D. J., Welcsh, P. L., Disteche, C. M., Filippova, G. N., Baliga, N. S., Aebersold, R., Ranish, J. A. & Krumm, A. (2008) CTCF physically links cohesin to chromatin, *Proceedings of the National Academy of Sciences of the United States of America.* 105, 8309-14.
76. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. & Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions, *Nature.* 485, 376-80.
77. Wang, H., Maurano, M. T., Qu, H., Varley, K. E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., Thurman, R. E., Kaul, R., Myers, R. M. & Stamatoyannopoulos, J. A. (2012) Widespread plasticity in CTCF occupancy linked to DNA methylation, *Genome research.* 22, 1680-8.
78. Sofueva, S., Yaffe, E., Chan, W. C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., Pollard, S. M., Schroth, G. P., Tanay, A. & Hadjur, S. (2013) Cohesin-mediated interactions organize chromosomal domain architecture, *The EMBO journal.* 32, 3119-29.
79. Zuin, J., Dixon, J. R., van der Reijden, M. I., Ye, Z., Kolovos, P., Brouwer, R. W., van de Corput, M. P., van de Werken, H. J., Knoch, T. A., van, I. W. F., Grosveld, F. G., Ren, B. & Wendt, K. S. (2014) Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells, *Proceedings of the National Academy of Sciences of the United States of America.* 111, 996-1001.
80. Collis, P., Antoniou, M. & Grosveld, F. (1990) Definition of the minimal requirements within the human beta-globin gene and the dominant control region for high level expression, *The EMBO journal.* 9, 233-40.
81. Tuan, D., Kong, S. & Hu, K. (1992) Transcription of the hypersensitive site HS2 enhancer in erythroid cells, *Proceedings of the National Academy of Sciences of the United States of America.* 89, 11219-23.
82. Hah, N., Danko, C. G., Core, L., Waterfall, J. J., Siepel, A., Lis, J. T. & Kraus, W. L. (2011) A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells, *Cell.* 145, 622-34.
83. Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A. Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., Oh, S., Kim, H. S., Glass, C. K. & Rosenfeld, M. G. (2013) Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation, *Nature.* 498, 516-20.

84. Lam, M. T., Cho, H., Lesch, H. P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., Benner, C., Kaikkonen, M. U., Kim, A. S., Kosaka, M., Lee, C. Y., Watt, A., Grossman, T. R., Rosenfeld, M. G., Evans, R. M. & Glass, C. K. (2013) Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription, *Nature.* 498, 511-5.
85. Melo, C. A., Drost, J., Wijchers, P. J., van de Werken, H., de Wit, E., Oude Vrielink, J. A., Elkon, R., Melo, S. A., Leveille, N., Kalluri, R., de Laat, W. & Agami, R. (2013) eRNAs are required for p53-dependent enhancer activity and gene transcription, *Molecular cell.* 49, 524-35.
86. Mousavi, K., Zare, H., Dell'orso, S., Grontved, L., Gutierrez-Cruz, G., Derfoul, A., Hager, G. L. & Sartorelli, V. (2013) eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci, *Molecular cell.* 51, 606-17.
87. Lin, Y. C., Benner, C., Mansson, R., Heinz, S., Miyazaki, K., Miyazaki, M., Chandra, V., Bossen, C., Glass, C. K. & Murre, C. (2012) Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate, *Nature immunology.* 13, 1196-204.
88. Hah, N., Murakami, S., Nagari, A., Danko, C. G. & Kraus, W. L. (2013) Enhancer transcripts mark active estrogen receptor binding sites, *Genome research.* 23, 1210-23.
89. Noonan, J. P. & McCallion, A. S. (2010) Genomics of long-range regulatory elements, *Annual review of genomics and human genetics.* 11, 1-23.
90. van Heyningen, V. & Williamson, K. A. (2002) PAX6 in sensory development, *Human molecular genetics.* 11, 1161-7.
91. Kleinjan, D. A., Seawright, A., Schedl, A., Quinlan, R. A., Danes, S. & van Heyningen, V. (2001) Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6, *Human molecular genetics.* 10, 2049-59.
92. de Kok, Y. J., Vossenaar, E. R., Cremers, C. W., Dahl, N., Laporte, J., Hu, L. J., Lacombe, D., Fischel-Ghodsian, N., Friedman, R. A., Parnes, L. S., Thorpe, P., Bitner-Glindzicz, M., Pander, H. J., Heilbronner, H., Graveline, J., den Dunnen, J. T., Brunner, H. G., Ropers, H. H. & Cremers, F. P. (1996) Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to the DFN3 gene POU3F4, *Human molecular genetics.* 5, 1229-35.
93. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A. & Visscher, P. M. (2009) Finding the missing heritability of complex diseases, *Nature.* 461, 747-53.
94. Visel, A., Rubin, E. M. & Pennacchio, L. A. (2009) Genomic views of distant-acting enhancers, *Nature.* 461, 199-205.
95. Emison, E. S., McCallion, A. S., Kashuk, C. S., Bush, R. T., Grice, E., Lin, S., Portnoy, M. E., Cutler, D. J., Green, E. D. & Chakravarti, A. (2005) A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk, *Nature.* 434, 857-63.
96. Rahimov, F., Marazita, M. L., Visel, A., Cooper, M. E., Hitchler, M. J., Rubini, M., Domann, F. E., Govil, M., Christensen, K., Bille, C., Melbye, M., Jugessur, A., Lie, R. T., Wilcox, A. J., Fitzpatrick, D. R., Green, E. D., Mossey, P. A., Little, J., Steegers-Theunissen, R. P., Pennacchio, L. A., Schutte, B. C. & Murray, J. C. (2008)

Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip, *Nature genetics.* 40, 1341-7.
97. Heinz, S., Romanoski, C. E., Benner, C., Allison, K. A., Kaikkonen, M. U., Orozco, L. D. & Glass, C. K. (2013) Effect of natural genetic variation on enhancer selection and function, *Nature.* 503, 487-92.
98. Boch, J. (2011) TALEs of genome targeting, *Nature biotechnology.* 29, 135-6.
99. Wang, H., Yang, H., Shivalila, C. S., Dawlaty, M. M., Cheng, A. W., Zhang, F. & Jaenisch, R. (2013) One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering, *Cell.* 153, 910-8.
100. Hong, J. W., Hendrix, D. A. & Levine, M. S. (2008) Shadow enhancers as a source of evolutionary novelty, *Science.* 321, 1314.