Genomic Views of Distant-Acting Enhancers

Axel Visel, Edward M. Rubin, and Len A. Pennacchio

Genomics Division, MS 84-171, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 USA.

U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598 USA.

Preface

In contrast to changes in protein-coding sequences, the significance of noncoding DNA variation in human disease has been minimally explored. A recent torrent of genome-wide association studies suggests that noncoding variation represents a significant risk factor for common disorders, but the mechanisms by which they contribute to disease remain largely obscure. As a major category of functional noncoding DNA, distant-acting transcriptional enhancers are likely involved in many developmental and disease-relevant processes. Genome-scale approaches for their discovery and functional characterization are now available and provide a growing knowledgebase for the systematic exploration of their role in human biology and disease susceptibility.

Introduction

Multiple lines of evidence indicate that important functional properties are embedded in the noncoding portion of the human genome, yet identifying and defining these features remains a major challenge. An initial glimpse of the magnitude of functional noncoding DNA was derived from comparative analysis of the first available mammalian genomes (human and mouse) which indicated that less than half of the evolutionarily constrained sequences in the human genome encode for proteins [1], a notion that was further reinforced when additional vertebrate genomes became available for comparative genomic analyses [2].

The overall impact of these presumably functional noncoding sequences on human biology was initially unclear. More recently, a considerable urgency to define their locations and functions came from a growing number of known associations of noncoding sequence variants with common human diseases. Specifically, genome-wide association studies (GWASs) have revealed a large number of disease susceptibility regions that do not overlap protein-coding genes, but rather map to noncoding intervals. As one of many examples, a 58kb linkage disequilibrium block located at human chromosome 9p21 was shown to be reproducibly associated with an increased risk for coronary artery disease, yet the risk interval is more than 60kb away from the nearest known protein-coding gene [3,4]. To estimate the global contribution of variation in noncoding sequences to phenotypic and disease traits, we performed a meta-analysis of ~1200 SNPs identified as the most significantly associated variants in GWASs published to date (www.genome.gov/26525384, accessed on March 2, 2009). Using conservative parameters that tend to overestimate the size of linkage disequilibrium blocks (details available upon request), we find that in 472 of 1170 (40%) cases no known exons overlap the linked SNP nor its associated haplotype block, suggesting that in more than a third of cases noncoding sequence variation causally contributes to the traits under investigation.

One noncoding function that could explain these GWAS hits are enhancers, a category of gene regulatory sequences that can act over long distances. A simplified view of our current understanding of the role of enhancers in regulating genes is summarized in Figure 1. The docking of RNA polymerase II to proximal promoter sequences and transcription initiation are fairly well characterized processes, whereas the mechanisms by which insulator and silencer elements buffer or repress gene regulation, respectively, are less well understood [5]. Transcriptional enhancers represent regulatory sequences that can be located upstream, downstream or within their target gene and can modulate expression independent of their orientation [6]. In vertebrates, enhancer sequences are thought to represent densely clustered aggregations of transcription factor binding sites [7]. When appropriate occupancy of transcription factor binding sites is achieved, recruitment of transcriptional co-activators and chromatin remodeling proteins occurs. The resulting protein aggregates are thought to facilitate DNA looping and ultimately promoter-mediated gene activation. Of note, in-depth studies of individual genes such as *APOE* or *NKX2-5* (reviewed in ref. 8) have shown that many genes are regulated by complex arrays of enhancers, each driving distinct aspects of the mRNA expression pattern. These modular properties of mammalian enhancers were also supported by their additive regulatory activities in heterologous recombination experiments [9].

The purely genetic evidence from GWASs does not allow any direct inferences regarding molecular mechanisms, but a number of in-depth studies of individual loci (see below) suggest that variation in distant-acting enhancer sequences and the resulting changes in their activities can contribute to human disorders. While we clearly expect a variety of other noncoding functional categories such as negative gene regulators or noncoding RNAs to play a role in human disease, in this review we will focus on the role of enhancers and on strategies to define their location and function genome-wide.

Main Text

Enhancers in Human Disease

Over the past half-century, beginning with the discovery that an inherited amino acid change in the beta-globin gene causes sickle-cell anemia [10,11], thousands of coding mutations in genes responsible for monogenic disorders were identified. In sharp contrast, the role of mutations not involving primary gene structural sequences has been minimally explored, largely due to the inability to recognize relevant noncoding sequences, let alone predict their function. Molecular genetic identification of individual enhancers involved in disease has been in most cases a painstaking and inefficient endeavor. Nevertheless, a number of successful studies have elegantly shown that distant-acting gene enhancers exist in the human genome and variation in their sequences can contribute to disease. In this section, we discuss several examples where enhancers were directly demonstrated to play a role in human disease: 1. thalassemias resulting from deletions or rearrangements of *beta-globin (HBB)* enhancers, 2. preaxial polydactyly resulting from *Sonic hedgehog (SHH)* limb enhancer point mutations, and 3. susceptibility to Hirschsprung disease associated with a *RET* proto-oncogene enhancer variant.

The extensive studies of the human globin system and its role in hemoglobinopathies have historically not only served as a test bed for defining the role of coding sequences in disease [10,11], but also for that of noncoding sequences. Alpha- and beta-thalassemias are hemoglobinopathies resulting from imbalances in the alpha- to beta-globin chain ratios in red blood cells. The molecular basis for these conditions was initially elucidated in those cases where inactivation or deletion of globin structural genes could be readily identified [12]. However, while gene deletion or sequence changes resulting in a truncated or nonfunctional gene product explained some thalassemia cases, for a subset of patients intensive sequencing efforts failed to reveal abnormalities in globin protein coding sequences. Through the extensive long-range mapping and sequencing of DNA

from individuals diagnosed with thalassemia but lacking globin coding mutations, it was eventually discovered that many of these globin chain imbalances were due to deletion or chromosomal rearrangements which resulted in the repositioning of distant-acting enhancers required for normal globin gene expression [13,14]. These early molecular genetic studies revealed a clear role for noncoding regulatory elements as a cause of human disorders through their impact on gene expression. Since then multiple such examples of "position effects", defined as a change in the expression of a gene when its location in a chromosome is changed, often by translocation, have been uncovered [15].

In addition to the pathological consequences of the removal or the repositioning of distant-acting enhancers, there are also examples of single nucleotide changes within enhancer elements as a cause of human disorders. One example of this category of disease-causing noncoding mutations involves the limb-specific ZRS (also known as MFCS1) long-distance enhancer of *Sonic hedgehog (SHH)* (Figure 2). This enhancer is located at the extreme distance of approximately one million base-pairs from *SHH* within the intron of a neighboring gene [16,17]. Of interest is the fact that initially the gene in which the enhancer resides was thought to be relevant for limb development based on mouse studies and was therefore named *limb region 1 (LMBR1)* [18]. Facilitated by the functional knowledge of the ZRS enhancer from mouse studies, targeted resequencing screens of this enhancer in humans revealed that it is associated with preaxial polydactyly. Approximately a dozen different single nucleotide variations in this regulatory element have been identified in humans with preaxial polydactyly and segregate with the limb abnormality in families [17,19]. Studies of the impact of the human ZRS sequence changes have been carried out in transgenic mice where the single nucleotide changes result in ectopic anterior limb expression during development, consistent with preaxial digit outgrowth [20]. Furthermore, sequence changes in the orthologous enhancers were found in mice as well as cats with preaxial polydactyly [21,22] and targeted deletion of the enhancer in mice causes

truncation of limbs [16]. These elegant studies illustrate the importance of first experimentally identifying distant acting enhancers to enable subsequent human genetic studies to explore the potential role of disease-causing mutation in functional noncoding sequences.

An additional example of enhancer variation contributing to human disease is the discovery of a common noncoding variant linked to disease susceptibility in Hirschsprung disease (HSCR). While multigenic, HSCR disease risk is strongly linked to coding mutations in the *RET* proto-oncogene [23,24]. However, familial studies have also revealed evidence for HSCR disease linked to the *RET* locus but lacking any accompanying functional *RET* coding mutations. Through the use of multi-species comparisons of orthologous genomic intervals including and flanking *RET* coupled with *in vitro* and *in vivo* functional studies, an enhancer sequence located in intron 1 of *RET* was identified and found to contain a common variant contributing greater than a 20-fold increased risk for HSCR disease compared to rarer alleles in this element [25,26]. Through mouse transgenesis experiments, this enhancer was shown to be active in the nervous system and digestive tract during embryogenesis in a way consistent with its putative role in HSCR [26]. It is interesting to note that while this enhancer variation is clearly important in disease risk, the variant alone is not sufficient to cause HSCR, highlighting the complex etiology of this disorder.

As is evident from these labor-intensive gene-centric studies, enhancers can in principle play an important role in disease, but it remains unclear whether they represent rare exceptions or if variation in enhancers contributes to disease on a pervasive scale. Support for the latter comes from a rapidly growing number of examples where noncoding SNPs linked to disease traits through GWASs were found to affect the expression levels of nearby genes [27], suggesting that variation in regulatory sequences may commonly contribute to a wide range of disorders. The results of the recent GWASs, coupled with the role of gene regulation in normal

human biology, provide a strong incentive for defining the distant-acting enhancer architecture of the human genome.

Harnessing Evolution

Gene-centric studies have been crucial for defining general characteristics of gene regulatory regions in specific human disorders but have only identified and characterized a limited number of such elements. Systematic large-scale identification of sequences that are likely to be enhancers was first enabled by comparative genomic strategies. These approaches are based on the assumption that the sequences of gene regulatory elements, like those of protein-coding genes, are under negative evolutionary selection because most changes in functional sequences have deleterious consequences [28-31]. Thus, it was hypothesized that statistical measures of evolutionary sequence constraint would provide a way to identify potential enhancer sequences within the vast amount of noncoding sequence in the human genome. Support for this approach initially came from retrospective comparative genomic analyses of experimentally well-defined enhancers revealing that they frequently shared sequence conservation with orthologous regions present in the genomes of other mammals. The observation that DNA conservation identified many of these complex regulatory elements encouraged investigators to transition from blind studies of regions flanking genes of interest to focusing specifically on noncoding sequences constrained across vertebrate species, culminating in whole-genome studies where conservation level alone drove experimentation [31-33].

Initially, comparisons across extreme evolutionary distances, such as between human and fish, were deemed most effective for this purpose [28,30]. Indeed, it was observed through large-scale transgenic mouse and fish studies that many of these noncoding sequences that had been conserved for hundreds of millions of years of evolution were enhancers that drove expression to highly specific anatomical structures during embryonic development. Likewise, so-called "ultraconserved"

noncoding elements which are blocks of 200bp or more that are perfectly conserved between human and rodents [34], were also found to be highly enriched in tissue-specific enhancers, suggesting that the success rate of comparative approaches for enhancer identification depends on scoring criteria, rather than just evolutionary distance [31]. This notion was further supported by the development of statistical tools specifically for this purpose, from which it became evident that even comparisons between relatively closely related species can be effective predictors of enhancers [2,35,36]. A large-scale transgenic mouse study that included nearly all non-exonic ultraconserved elements in the human genome revealed that while many of them are developmental *in vivo* enhancers, other noncoding conserved sequences that are under similar evolutionary constraint, but less than perfectly conserved between human and rodents, are equally enriched in enhancers [32]. These results suggest that ultraconserved elements do not represent a functionally distinct subgroup of conserved noncoding sequences regarding their enrichment for *in vivo* enhancers, but rather that there is a much larger number of noncoding sequences that are under similar evolutionary constraint and just as enriched in enhancers as ultraconserved elements.

Independent of the specific algorithms and metrics that were used, most categories of conserved noncoding sequences were found to be not randomly distributed in the genome. Instead, they are located in a highly biased manner near genes active during development [2,32-34], consistent with the observation that a large fraction of these noncoding sequences give robust positive signals in various assays as tissue-specific *in vivo* enhancers active during development.

Comparative approaches are an effective high throughput genomic strategy for identifying noncoding sequences with a high likelihood of being an enhancer, but they suffer from several limitations. First, while conservation is indicative of function, it is not necessarily indicative of enhancer activity because many other types of noncoding functional elements are known to exist that may have similar conservation signatures. Second, even when conserved noncoding DNA is due to

enhancer function, conservation cannot predict when and where an enhancer is active in the developing or adult organism. For all identified candidates experimental studies are needed to decipher the gene regulatory properties of each element and these transgenic studies cannot feasibly be scaled to generate truly comprehensive genome-wide datasets.

A perplexing study questioning the importance of extremely conserved enhancers was the lack of an apparent phenotype upon targeted deletion of four independent ultraconserved elements in mice [37]. General expectations were that noncoding sequences perfectly conserved in mammals for dozens of millions of years must be essential and their deletion should result in severe phenotypes, comparable to those observed upon deletion of the *Shh* limb enhancer and other less well-conserved enhancers [8,16]. However, mice with deletions of such ultraconserved enhancers were viable, fertile and showed no overt phenotype [37]. Interpretations of this lack of obvious effects are similar to those for absence of phenotypes upon deletion of highly conserved protein-coding genes: Minor phenotypes may have escaped detection in the assays used, functional redundancy with other genes or enhancers, or reductions in fitness that only become apparent over multiple generations or are not easily detected in a controlled laboratory environment. This study highlighted that while extreme noncoding sequence conservation is an effective predictor of the location of enhancers in the genome, the degree of evolutionary constraint is not directly correlated with the severity of anticipated phenotypes.

Sequencing-Based Enhancer Discovery

As a complementary strategy to comparative genomic methods, it has recently become possible to generate genome-wide maps of chromatin marks that can be used to identify the location of enhancers and other regulatory regions. These genomic approaches have been enabled by (a) an improved understanding of the proteins and epigenetic marks found at particular categories of regulatory

elements and (b) concurrently developed technologies that allow traditional chromatin-immunoprecipitation techniques to be applied on the scale of whole vertebrate genomes. In particular, the initial in-depth studies of 1% of the genome in the ENCODE pilot project, largely based on datasets generated by the ChIP-chip technique (see Text Box 1), revealed molecular properties of a variety of regulatory elements. With respect to enhancer identification, a particularly relevant insight was the identification of specific methylation signatures found at enhancers. In contrast to promoters, which are marked by trimethylation of histone H3 at lysine residue 4 (H3K4me3), active enhancers are marked by monomethylation (H3K4me1) at this position [38]. Mapping these marks in the ENCODE regions and, more recently, throughout the entire genome [39] revealed tens of thousands of elements that were predicted to be active enhancers in the examined cell types. Importantly, these predicted enhancers were also frequently associated with the transcriptional coactivators p300 and/or TRAP220, raising the possibility that such coactivators might represent useful general markers for mapping enhancers. While it was initially not clear to what extent the presence of transcriptional coactivators like p300 is indicative of active vs. inactive enhancers, comparison of DNase I hypersensitivity (DNaseI HS, a marker of open chromatin structure) in several cell lines throughout the ENCODE regions revealed that the location of cell line-specific distal DNaseI HS sites correlates with cell line-specific p300 binding at these sites, providing further support for the possibility that transcriptional coactivators, along with histone modification signatures, may be useful for mapping of DNA elements with cell- and tissue-specific enhancer activities [40].

Through the introduction of the ChIP-seq technique (see Text Box 2), which has now superseded ChIP-chip as the method of choice for many applications, genome-wide maps for a considerable number of chromatin marks and transcription factors both in human and mouse have become available [41-53]. In addition to the H3K4me1/3 signature discussed above, these datasets enabled the identification of additional chromatin marks present at predicted or validated

enhancers and provided a refined view of their correlation to enhancer activities [42,49,53]. However, with very few exceptions (e.g., references 48,52) genome-wide mapping of these and other regulation-associated chromatin marks (see Table 1) was done in immortalized cell lines, cultured stem cells or primary cell cultures. Thus, the maps of potentially enhancer-associated marks produced through these studies provided limited insight into their *in vivo* distribution during embryonic development and in adult organs, likely concealing the genomic location of enhancers that are inactive in these cells.

In a recent ChIP-seq study targeted at the prediction of enhancers that are active in a particular tissue during embryonic development, the transcriptional coactivator p300 was mapped in chromatin directly derived from embryonic mouse tissues including the forebrain, the midbrain, and the limb buds [54]. Overall, several thousand p300 peaks were identified from these three tissues, with the vast majority of genome regions only being significantly enriched in one of the three tissues and located in noncoding regions distal from known promoters. Transgenic mouse experiments with close to a hundred of these sequences revealed that they are in almost all cases developmental enhancers. More importantly, the tissue-specific occupancy by p300 as identified by ChIP-seq could in most cases also accurately predict the *in vivo* patterns of expression driven by these enhancers, providing an important advantage over comparative genomic methods for enhancer identification. It was also shown that tissue-specific p300 peaks are globally enriched near genes that are expressed in the same tissue, again consistent with their hypothesized function as active transcriptional enhancers.

These experimentally predicted genome-wide sets of *in vivo* enhancers also made it possible to address the controversial issue to what extent evolutionary conservation is a hallmark of *in vivo* enhancers [55]. Several studies have shown that highly conserved noncoding elements are enriched in developmental *in vivo* enhancers [31-33]. However, some observations have challenged such a generalized correlation between sequence conservation and enhancer activity: (1) experimental

analysis of individual loci suggested that a large proportion of enhancers cannot be detected by comparative genomics [56], (2) a surprisingly large fraction of sequences in the ENCODE regions whose molecular marks suggest regulatory functions were not or only weakly conserved [57], (3) histone methylations present at orthologous loci in human and mouse did not correlate with overall increased levels of sequence conservation [58]. In contrast to these findings, approximately 90% of the tissue-specific p300 peaks identified by ChIP-seq in developing mouse tissues overlapped regions that are under detectable evolutionary constraint [54]. While there may be variation in the degree of evolutionary constraint of enhancers that are active in different types of cells or developing tissues, these data suggest that developmental enhancers that can be identified through p300-binding are commonly constrained.

While in its infancy, the selected studies reviewed here highlight the clear potential of mapping various chromatin marks for identifying and predicting the activity of transcriptional enhancers on a genome-wide scale. The continued progress in throughput and cost reductions of next-generation sequencing technologies offers an increasingly powerful genome-wide means for identifying specific DNA-protein interactions. Spurred by continued improvements in sequencing, we anticipate that high-resolution genome-wide *in vivo* maps of chromatin marks will become available for comprehensive series of developing and adult tissues in normal as well as disease states, providing multi-layered *in vivo* annotations of the noncoding portion of our genome. It is important that we realize that despite this expected progress, we will continue to need parallel *in vitro* and *in vivo* biological studies to understand the functions associated with chromatin marks and to conclusively study the mechanisms by which sequence variation in distant-acting enhancers contributes to disease.

Defining the Targets

The methods described above have considerably improved our capability to identify enhancers and their associated activity patterns on a genomic scale, but a remaining important challenge will be to determine the relations between enhancers and genes. Currently, comparing ChIP-chip or ChIP-seq data with transcriptome data from microarrays or RNA-seq [59] can provide highly suggestive clues what the target gene of a given enhancer in a given tissue is, but does not provide the direct evidence for enhancer-promoter interactions that would be desirable to map tissue-specific regulatory networks on a genomic scale.

Early circumstantial evidence suggested that long-distance regulation of genes by enhancers occurs through the formation of physical chromatin loops, yet it became first possible to study such interactions systematically through the introduction of the chromosome conformation capture (3C) assay and its derivative technologies [60]. Similar to ChIP, the 3C approach relies on formaldehyde cross-linking to capture DNA-DNA interactions directly in intact cells or cell nuclei. Previously hypothesized pairs of interacting sites are then tested and validated in a one-by-one fashion through the quantitation of cross-linking events. As one of many examples demonstrating the utility of 3C in the analysis of distant-acting vertebrate enhancers, Amano et al. recently used this technique to study chromatin interactions at the *Shh* hedgehog locus whose role in limb development we discussed in detail above [61]. Using the 3C technique, the authors demonstrated elegantly that the limb-specific long-range enhancer located in an intron of the *Lmbr1* gene directly interacts with increased frequency with the *Shh* promoter in limb buds but not in other tissues tested, providing important mechanistic support for its proposed role in *Shh* gene regulation in limb development. As an alternative approach to 3C, RNA-tagging and recovery of associated proteins (RNA-TRAP) can also be used to establish physical proximity between distal noncoding sequences and actively transcribed genes, which was first demonstrated in the mouse *beta-globin* locus [62].

This work and other gene-centric studies (for more examples see references 63,64) were critical to shape our understanding of enhancer-promoter interactions. However, they suffer from the fundamental limitation that only one or very few previously hypothesized interactions between specific loci can be assayed per experiment. This limitation was partially overcome through the use of microarrays to analyze entire 3C libraries (chromosome conformation capture-on-chip, 4C [65] and circular chromosome conformation capture, also called 4C [66]). By applying this approach to fetal liver and brain, it was demonstrated that the *beta-globin* locus control region (LCR) makes reproducible tissue-specific contacts with other loci predominantly located on the same chromosome, yet in some cases dozens of megabases away from the LCR [65]. Of possible relevance for adopting this approach for enhancer discovery, reproducible interactions with other chromosome regions were also observed in the brain where the LCR is thought to be inactive.

The 4C approaches described above represented a significant improvement, but they still preclude the generation of truly genome-wide interaction networks because each experiment only reveals the genome-wide interactions of a single site of interest. This problem is partially alleviated by the chromosome conformation capture carbon copy (5C) method [67] in which a complex 3C library generated through multiplexed PCR is analyzed by large-scale sequencing to generate a comprehensive "many-to-many" interaction map of DNA-DNA-interactions. However, due to the need for specific primers for each possible interacting fragment and the sequencing depth required for analysis of the resulting libraries, application of 5C has so far been restricted to the in-depth analysis of single loci or chromosome regions.

As an alternative genome-wide approach, it has been proposed to use antibody-based methods rather than selecting particular chromosome regions to restrict the analysis space for studying DNA-DNA interactions to a size that can be affordably analyzed by currently available sequencing technologies. Namely, it

was suggested to couple a chromatin interaction paired-end tag sequencing (ChIA-PET) strategy to a ChIP step that enriches for chromatin fragments bound to a specific transcription factor or other chromatin mark of interest [63]. While the technical feasibility of this approach remains to be demonstrated, it has remarkable potential for enhancer discovery as its application to general enhancer-associated marks like p300 or histone methylations [38,54] might in a single step identify enhancers active in a tissue of interest, as well as their respective target genes.

Perspective

In the past, genetic and medical resequencing studies have been empowered by knowledge about the structure of protein-coding genes and a detailed understanding of the relation between mRNA sequences and the primary structure of the proteins they encode. Through such studies, disease links have been established for a sizeable fraction of the ~20,000 protein-encoding genes. In contrast, a very limited number of sequence changes in gene regulatory sequences could be linked to human disease. Consequently, an important impetus for functionally annotating the noncoding portion of the human genome and the *cis*-regulatory elements it contains is to assess the relationship between variations in noncoding sequences and human disease. In the absence of genome-wide catalogues of functionally annotated regulatory elements, their impact on human biology as well as disease will remain an untested hypothesis.

In this review, we have outlined how the number of annotated noncoding regulatory sequences is poised to dramatically expand through the continued progress of DNA sequencing technologies coupled with markers to assess higher order chromatin status. Nevertheless, functionally characterizing the distant-acting enhancer architecture of the human genome in its entirety will be an enormous undertaking due to the vast number of data points needed, which

include dozens of tissues and cell types, as well as developmental and possibly disease states.

A further unmet challenge in large-scale studies of distant-acting enhancers will be to link them to the genes they regulate. Again, it is expected that considerable progress towards identification of genome-wide interaction maps will be achieved through methods derived from existing strategies for chromatin conformation analysis coupled to advanced sequencing technologies. Such linking enhancers to their cognate gene will allow the further binning of these functional sequences into their basic "gene" unit of heredity for collective resequencing analysis.

In conclusion, it is important to keep in mind that several categories of functional elements in addition to enhancers exist in the noncoding genome (e.g. insulators, negative regulators, promoters, and non-protein-encoding RNA functions). While this review focused on distant-acting enhancers, these other types of noncoding sequences will also be crucial targets for large-scale identification and characterization and it is expected that technologies similar to those described here for enhancer identification will make it possible to also explore their role in human biology and disease.

Acknowledgments

The authors wish to thank M. Blow, S. Deutsch, and A. Sczyrba for help with computational analysis of GWAS data and C. Attanasio for critical comments. L.A.P./E.M.R. were supported by the Berkeley-PGA, under the Programs for Genomic Applications, funded by National Heart, Lung, & Blood Institute, and L.A.P. by the National Human Genome Research Institute. This work was supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Table 1. Selected major categories of noncoding functional elements.

Category	Function	Selected Associated Chromatin Marks*
Promoter	Region located immediately 5' of a protein-encoding gene that binds to RNA polymerase II and from which transcription is initiated	PolII [42] H3K4me3 [38] (active promoters)
Enhancer	Region that activates transcription, often in a temporally and spatially restricted manner, by acting on a promoter. Enhancers can be located far away from target promoters and are orientation-independent	p300 [38,54] H3K4me1 [38]
Insulator	Separates active from inactive chromatin domains and interferes with enhancer activity when placed between and enhancer and promoter	CTCF [42,51]
Repressor/Silencer	Negative regulators of gene expression	REST [43] Suz12 [68,69]

* Many additional chromatin marks were found to correlate with one or several of these categories of regulatory elements. Detailed descriptions of these markers and their respective binding characteristics at different types of regulatory sequence elements can be found in references 38,39,42,49,53.

Text Box 1: Mapping of Regulatory Elements by ChIP-chip

Formaldehyde cross-linking of DNA to proteins that bind to it directly or as part of larger complexes [70] combined with subsequent immunoprecipitation targeting specific DNA-associated proteins (ChIP, [71]) has been widely used in the pre-genomic era to study protein-DNA interactions directly in living cells. The technique involves the molecular fixation of non-covalent protein-DNA interactions, shearing of the cross-linked chromatin, immunoprecipitation with an antibody binding the protein (or protein modification) of interest, and subsequent quantitation of enrichment of the DNA fragments associated with the targeted protein of interest compared to non-immunoprecipitated ("input") DNA. While useful to examine protein-DNA interactions at individual hypothesized binding locations, the need for quantitation at every single site of interest initially thwarted the application of this technique on a genomic scale. However, this changed dramatically with the introduction of DNA microarrays that enabled hybridization-based quantitation of large numbers of candidate sites in parallel ("ChIP-on-chip" or "ChIP-chip"), thus making it possible to screen in a single experiment entire compact model organism genomes [72,73] or large vertebrate genome intervals [74] to identify all binding sites of a given protein in a cell type or tissue of interest (Figure 3). This technique was used on a massive scale by the ambitious multi-center Encyclopedia of DNA Elements (ENCODE) pilot project, where dozens of proteins and protein modifications were initially mapped in a representative 1% portion of the human genome [57].

Text Box 2: Mapping of Regulatory Elements by ChIP-seq

Recently, chromatin immunoprecipitation coupled to massively-parallel sequencing (ChIP-seq) has become increasingly utilized as an alternative to ChIP-chip for mapping regulatory elements in the genome [42-45]. The ChIP-seq method is very similar to the experimental setup of ChIP-chip, except that in the final step, massive-parallel sequencing techniques are used to determine the sequence of immunoprecipitated DNA fragments, which are then computationally mapped to the reference genome. Identification of chromosome regions that are significantly enriched in reads then reveals the genomic locations of the chromatin mark of interest (Figure 3). Advantages of the ChIP-seq method compared to ChIP-chip include: 1. Through next-generation sequencing technologies it is now possible to obtain millions of mappable reads in a single experiment at moderate cost. 2. The results from ChIP-seq are based on statistical analysis of read counts, which overcomes many of the challenges associated with the quantitation and normalization of hybridization signals, especially as an increasing number of advanced computational ChIP-seq analysis tools are becoming available [75]. 3. The analysis covers by default the entire mappable portion of the reference genome without the need to restrict the analysis to subregions of the genome.

Figure 1. Schematic overview of gene regulation by distant-acting enhancers. For many genes, the regulatory information embedded in the promoter is insufficient to drive the complex expression pattern observed at the mRNA level, suggesting that appropriate expression in time and space depends on additional distant-acting *cis*-regulatory sequences. Tissue-specific enhancers are thought to contain combinations of binding sites for different transcription factors. Only when all required transcription factors are present in a tissue, the enhancer becomes active: it binds to transcriptional co-activators, relocates into physical proximity of the gene promoter through a looping mechanism, and activates transcription by RNA polymerase II. In any given tissue, only a subset of enhancers is active, as shown here schematically for two separate enhancers with brain- and limb-specific activities. Insulator elements prevent enhancer-promoter interactions and can thus restrict the activity of enhancers to defined chromatin domains. In addition to activation by enhancers, negative regulatory elements including repressors and silencers can contribute to transcriptional regulation (not shown).

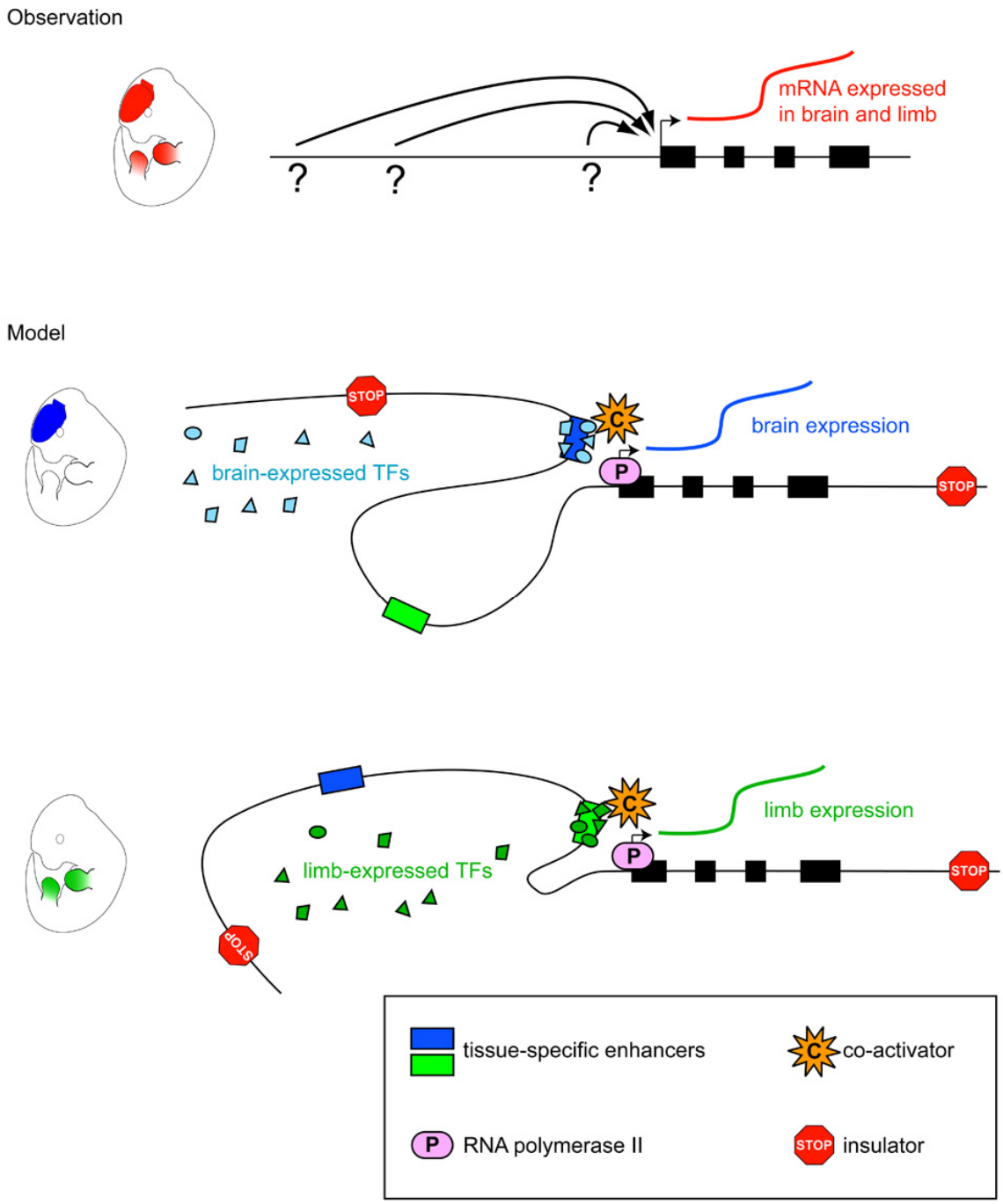

Observation
mRNA expressed
in brain and limb
?
?
?
Model
STOP
brain-expressed TFs
C
P
brain expression
STOP
limb-expressed TFs
C
P
limb expression
STOP
STOP
tissue-specific enhancers
co-activator
RNA polymerase II
insulator

Figure 2. Consequences of deletion and mutation of the limb enhancer of *Sonic hedgehog*. A) The limb enhancer of the *Sonic hedgehog* gene is located approximately 1 megabase away from its target promoter in the intron of a neighboring gene (*Lmbr1,* exons not shown). In transgenic mouse reporter assays, this noncoding sequence targets gene expression to a posterior region of the developing limb bud [17]. B) Mice with a targeted deletion of this enhancer have severely truncated limbs, which strikingly demonstrates its functional importance in development [16]. C-E) Several point mutations in the orthologous human enhancer sequence were shown to result in preaxial polydactyly, emphasizing the potential significance of variation in noncoding functional sequences in both rare and common human disorders [17]. C, D) Hands of two different patients with point mutations in the *Sonic hedgehog* limb enhancer. E) Point mutations associated with preaxial polydactyly identified in four unrelated families.

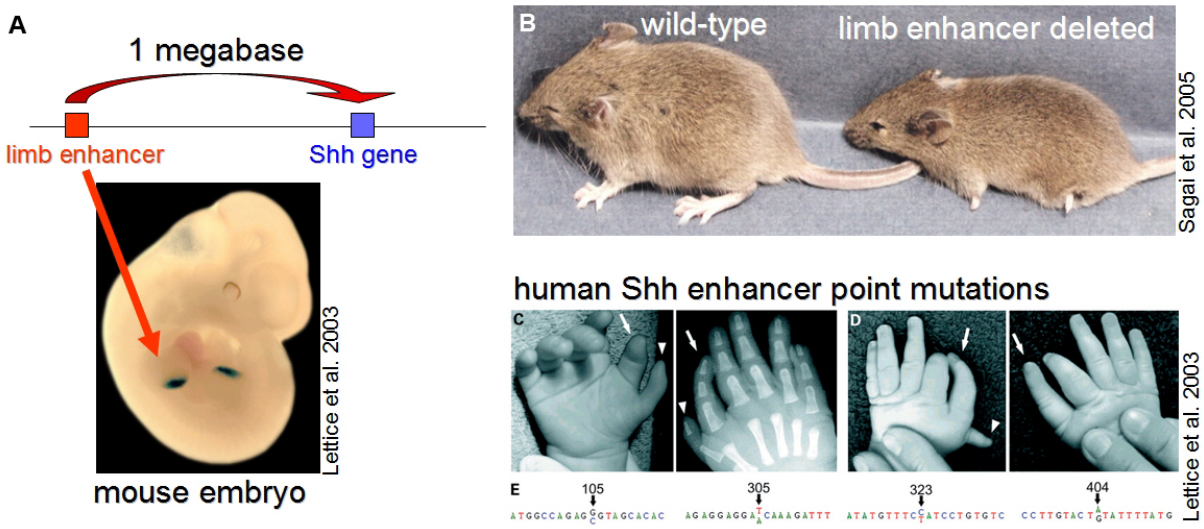
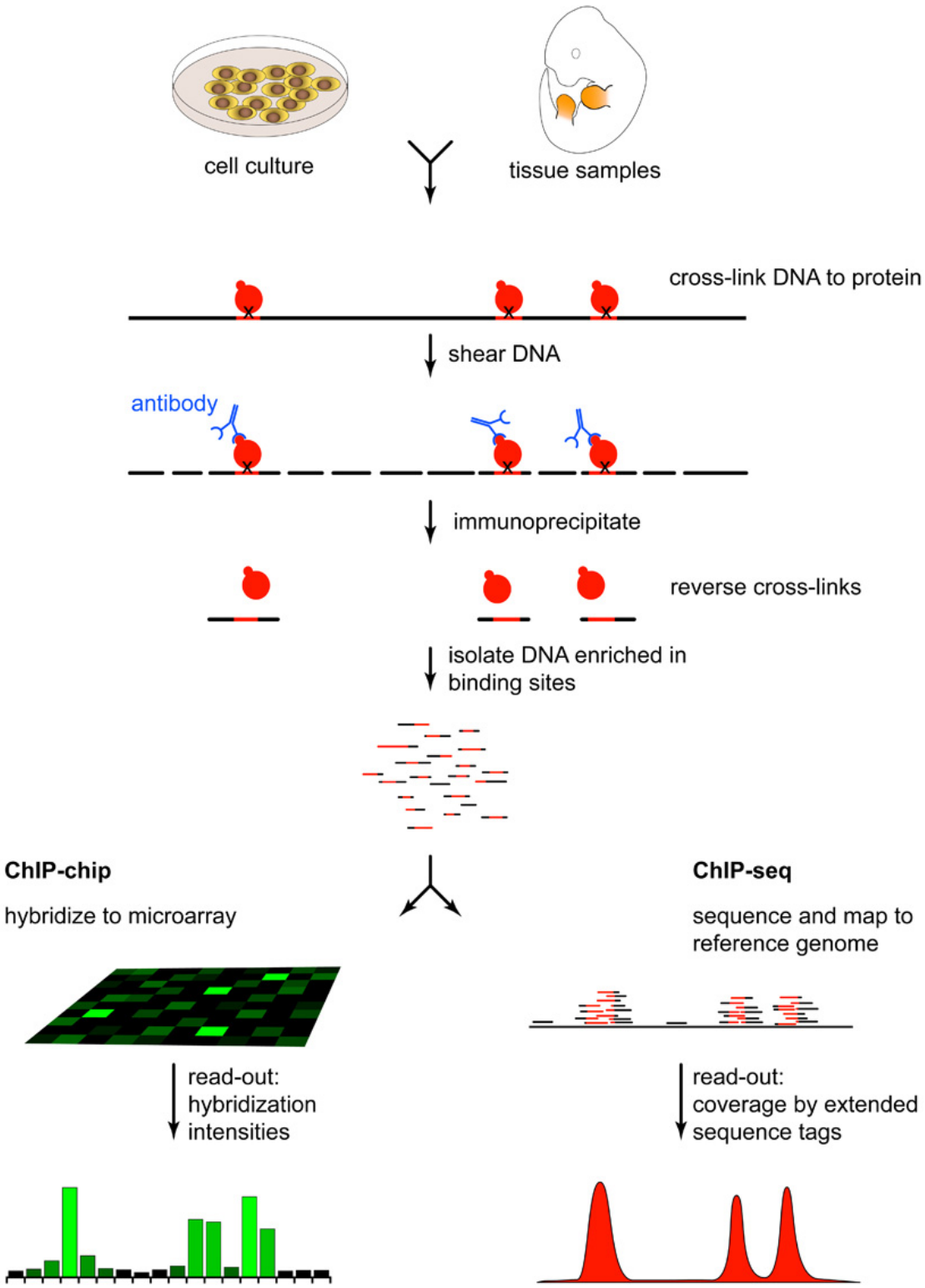


Figure 3. Schematic overview of ChIP-chip and ChIP-seq methods. Both approaches depend on the cross-linking of protein to DNA by formaldehyde, either in cultured cells or in tissue samples. After cross-linking, chromatin is sheared and a suitable antibody is used to enrich for DNA fragments bound to a protein of interest. In many cases, antibodies binding to covalently modified proteins are used, e.g. recognizing methyl groups at defined amino acid residues of histones. Following immunoprecipitation and reversal of cross-links, the DNA libraries enriched in binding sites for the chromatin mark of interest are either analyzed by hybridization to microarrays (ChIP-chip) or by massively-parallel sequencing and alignment of the obtained sequence reads to the reference genome (ChIP-seq). See Text Boxes 1 and 2 for additional details about these techniques.


cell culture
tissue samples
cross-link DNA to protein
shear DNA
antibody
immunoprecipitate
reverse cross-links
isolate DNA enriched in binding sites
ChIP-chip
hybridize to microarray
read-out: hybridization intensities
ChIP-seq
sequence and map to reference genome
read-out: coverage by extended sequence tags

References

References

[1] Waterston, R.H. *et al.*, Initial sequencing and comparative analysis of the mouse genome. *Nature* 420 (6915), 520-562 (2002).

[2] Siepel, A. *et al.*, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15 (8), 1034-1050 (2005).

[3] Helgadottir, A. *et al.*, A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* 316 (5830), 1491-1493 (2007).

[4] McPherson, R. *et al.*, A common allele on chromosome 9 associated with coronary heart disease. *Science* 316 (5830), 1488-1491 (2007).

[5] Maston, G.A., Evans, S.K., & Green, M.R., Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 7, 29-59 (2006).

[6] Banerji, J., Rusconi, S., & Schaffner, W., Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27 (2 Pt 1), 299-308 (1981).

[7] Panne, D., The enhanceosome. *Curr Opin Struct Biol* 18 (2), 236-242 (2008).

[8] Visel, A., Bristow, J., & Pennacchio, L.A., Enhancer identification through comparative genomics. *Semin Cell Dev Biol* 18 (1), 140-152 (2007).

[9] Visel, A. *et al.*, Functional autonomy of distant-acting human enhancers. *Genomics* 93 (6), 509-513 (2009).

[10] Ingram, V.M., Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature* 180 (4581), 326-328 (1957).

[11] Pauling, L., Itano, H.A., & et al., Sickle cell anemia a molecular disease. *Science* 110 (2865), 543-548 (1949).

[12] Kan, Y.W. *et al.*, Deletion of alpha-globin genes in haemoglobin-H disease demonstrates multiple alpha-globin structural loci. *Nature* 255 (5505), 255-256 (1975).

[13] Kioussis, D., Vanin, E., deLange, T., Flavell, R.A., & Grosveld, F.G., Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. *Nature* 306 (5944), 662-666 (1983).

[14] Semenza, G.L. *et al.*, The silent carrier allele: beta thalassemia without a mutation in the beta-globin gene or its immediate flanking regions. *Cell* 39 (1), 123-128 (1984).

[15] Kleinjan, D.A. & Lettice, L.A., Chapter 13 long-range gene control and genetic disease. *Adv Genet* 61, 339-388 (2008).

[16] Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., & Shiroishi, T., Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* 132 (4), 797-803 (2005).

[17] Lettice, L.A. *et al.*, A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* 12 (14), 1725-1735 (2003).

[18] Clark, R.M., Marker, P.C., & Kingsley, D.M., A novel candidate gene for mouse and human preaxial polydactyly with altered expression in limbs of Hemimelic extra-toes mutant mice. *Genomics* 67 (1), 19-27 (2000).

[19] Furniss, D. *et al.*, A variant in the sonic hedgehog regulatory sequence (ZRS) is associated with triphalangeal thumb and deregulates expression in the developing limb. *Hum Mol Genet* 17 (16), 2417-2423 (2008).

[20] Masuya, H. *et al.*, A series of ENU-induced single-base substitutions in a long-range cis-element altering Sonic hedgehog expression in the developing mouse limb bud. *Genomics* 89 (2), 207-214 (2007).

[21] Lettice, L.A., Hill, A.E., Devenney, P.S., & Hill, R.E., Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. *Hum Mol Genet* 17 (7), 978-985 (2008).

[22] Lettice, L.A. *et al.*, Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc Natl Acad Sci U S A* 99 (11), 7548-7553 (2002).

[23] Bolk, S. *et al.*, A human model for multigenic inheritance: phenotypic expression in Hirschsprung disease requires both the RET gene and a new 9q31 locus. *Proc Natl Acad Sci U S A* 97 (1), 268-273 (2000).

[24] Gabriel, S.B. *et al.*, Segregation at three loci explains familial and population risk in Hirschsprung disease. *Nat Genet* 31 (1), 89-93 (2002).

[25] Emison, E.S. *et al.*, A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature* 434 (7035), 857-863 (2005).

[26] Grice, E.A., Rochelle, E.S., Green, E.D., Chakravarti, A., & McCallion, A.S., Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer. *Hum Mol Genet* 14 (24), 3837-3845 (2005).

[27] Cookson, W., Liang, L., Abecasis, G., Moffatt, M., & Lathrop, M., Mapping complex disease traits with global gene expression. *Nat Rev Genet* 10 (3), 184-194 (2009).

[28] Aparicio, S. *et al.*, Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. *Proc Natl Acad Sci U S A* 92 (5), 1684-1688 (1995).

[29] Loots, G.G. *et al.*, Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 288 (5463), 136-140. (2000).

[30] Nobrega, M.A., Ovcharenko, I., Afzal, V., & Rubin, E.M., Scanning human gene deserts for long-range enhancers. *Science* 302 (5644), 413 (2003).

[31] Pennacchio, L.A. *et al.*, In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444 (7118), 499-502 (2006).

[32] Visel, A. *et al.*, Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* 40 (2), 158-160 (2008).

[33] Woolfe, A. *et al.*, Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* 3 (1), e7 (2005).

[34] Bejerano, G. *et al.*, Ultraconserved elements in the human genome. *Science* 304 (5675), 1321-1325 (2004).

[35] Prabhakar, S. *et al.*, Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res* 16 (7), 855-863 (2006).

[36] Cooper, G.M. *et al.*, Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 15 (7), 901-913 (2005).

[37] Ahituv, N. *et al.*, Deletion of Ultraconserved Elements Yields Viable Mice. *PLoS Biol* 5 (9), e234 (2007).

[38] Heintzman, N.D. *et al.*, Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 39 (3), 311-318 (2007).

[39] Heintzman, N.D. *et al.*, Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459 (7243), 108-112 (2009).

[40] Xi, H. *et al.*, Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet* 3 (8), e136 (2007).

[41] Wei, C.L. *et al.*, A global map of p53 transcription-factor binding sites in the human genome. *Cell* 124 (1), 207-219 (2006).

[42] Barski, A. *et al.*, High-resolution profiling of histone methylations in the human genome. *Cell* 129 (4), 823-837 (2007).

[43] Johnson, D.S., Mortazavi, A., Myers, R.M., & Wold, B., Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316 (5830), 1497-1502 (2007).

[44] Robertson, G. *et al.*, Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 4 (8), 651-657 (2007).

[45] Mikkelsen, T.S. *et al.*, Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448 (7153), 553-560 (2007).

[46] Zhao, X.D. *et al.*, Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell* 1 (3), 286-298 (2007).

[47] Chen, X. *et al.*, Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133 (6), 1106-1117 (2008).

[48] Wederell, E.D. *et al.*, Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res* 36 (14), 4549-4564 (2008).

[49] Robertson, A.G. *et al.*, Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res* 18 (12), 1906-1917 (2008).

[50] Ku, M. *et al.*, Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet* 4 (10), e1000242 (2008).

[51] Cuddapah, S. *et al.*, Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* 19 (1), 24-32 (2009).

[52] Gao, N. *et al.*, Dynamic regulation of Pdx1 enhancers by Foxa1 and Foxa2 is essential for pancreas development. *Genes Dev* 22 (24), 3435-3448 (2008).

[53] Wang, Z. *et al.*, Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* 40 (7), 897-903 (2008).

[54] Visel, A. *et al.*, ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457 (7231), 854-858 (2009).

[55] Cooper, G.M. & Brown, C.D., Qualifying the relationship between sequence conservation and molecular function. *Genome Res* 18 (2), 201-205 (2008).

[56] McGaughey, D.M. *et al.*, Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. *Genome Res* 18 (2), 252-260 (2008).

[57] ENCODE Project Consortium *et al.*, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447 (7146), 799-816 (2007).

[58] Bernstein, B.E. *et al.*, Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* 120 (2), 169-181 (2005).

[59] Wang, Z., Gerstein, M., & Snyder, M., RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10 (1), 57-63 (2009).

[60] Dekker, J., Rippe, K., Dekker, M., & Kleckner, N., Capturing chromosome conformation. *Science* 295 (5558), 1306-1311 (2002).

[61] Amano, T. *et al.*, Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev Cell* 16 (1), 47-57 (2009).

[62] Carter, D., Chakalova, L., Osborne, C.S., Dai, Y.F., & Fraser, P., Long-range chromatin regulatory interactions in vivo. *Nat Genet* 32 (4), 623-626 (2002).

[63] Fullwood, M.J., Wei, C.L., Liu, E.T., & Ruan, Y., Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res* 19 (4), 521-532 (2009).

[64] Miele, A. & Dekker, J., Long-range chromosomal interactions and gene regulation. *Mol Biosyst* 4 (11), 1046-1057 (2008).

[65] Simonis, M. *et al.*, Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* 38 (11), 1348-1354 (2006).

[66] Zhao, Z. *et al.*, Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* 38 (11), 1341-1347 (2006).

[67] Dostie, J. *et al.*, Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 16 (10), 1299-1309 (2006).

[68] Lee, T.I. *et al.*, Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* 125 (2), 301-313 (2006).

[69] Squazzo, S.L. *et al.*, Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res* 16 (7), 890-900 (2006).

[70] Van Lente, F., Jackson, J.F., & Weintraub, H., Identification of specific crosslinked histones after treatment of chromatin with formaldehyde. *Cell* 5 (1), 45-50 (1975).

[71] Solomon, M.J., Larsen, P.L., & Varshavsky, A., Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* 53 (6), 937-947 (1988).

[72] Ren, B. *et al.*, Genome-wide location and function of DNA binding proteins. *Science* 290 (5500), 2306-2309 (2000).

[73] Iyer, V.R. *et al.*, Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409 (6819), 533-538 (2001).

[74] Horak, C.E. *et al.*, GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. *Proc Natl Acad Sci U S A* 99 (5), 2924-2929 (2002).

[75] Barski, A. & Zhao, K., Genomic location analysis by ChIP-Seq. *J Cell Biochem* 107 (1), 11-18 (2009).