Genomic views of distant-acting enhancers

Axel Visel[1,2], Edward M. Rubin[1,2] & Len A. Pennacchio[1,2]

In contrast to protein-coding sequences, the significance of variation in non-coding DNA in human disease has been minimally explored. A great number of recent genome-wide association studies suggest that non-coding variation is a significant risk factor for common disorders, but the mechanisms by which this variation contributes to disease remain largely obscure. Distant-acting transcriptional enhancers — a major category of functional non-coding DNA — are involved in many developmental and disease-relevant processes. Genome-wide approaches to their discovery and functional characterization are now available and provide a growing knowledge base for the systematic exploration of their role in human biology and disease susceptibility.

Multiple lines of evidence indicate that important functional properties are embedded in the non-coding portion of the human genome, but identifying and defining these features remains a major challenge. An initial estimate of the magnitude of functional non-coding DNA was derived from comparative analysis of the first available mammalian genomes (human and mouse), which indicated that fewer than half of the evolutionary constrained sequences in the human genome encode proteins[1], a prospect that gained further support when additional vertebrate genomes became available for comparative genomic analyses[2].

The overall impact of these presumably functional non-coding sequences on human biology was initially unclear. A considerable urgency to define their locations and functions came from a growing number of known associations of non-coding sequence variants with common human diseases. Specifically, genome-wide association studies (GWAS) have revealed a large number of disease susceptibility regions that do not overlap protein-coding genes but rather map to non-coding intervals. For example, a 58-kilobase 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 lies more than 60 kilobases away from the nearest known protein-coding gene[3,4]. To estimate the global contribution of variation in non-coding sequences to phenotypic and disease traits, we performed a meta-analysis of ~1,200 single-nucleotide polymorphisms (SNPs) identified as the most significantly associated variants in GWAS published so far (ref. 5, accessed 2 March 2009). Using conservative parameters that tend to overestimate the size of linkage disequilibrium blocks, we found that in 40% of cases (472 of 1,170) no known exons overlap either the linked SNP or its associated haplotype block, suggesting that in more than one-third of cases non-coding sequence variation causally contributes to the traits under investigation.

One possibility that could explain these GWAS hits is that the non-coding intervals contain enhancers, a category of gene regulatory sequence that can act over long distances. A simplified view of the current understanding of the role of enhancers in regulating genes is summarized in Fig. 1. The docking of RNA polymerase II to proximal promoter sequences and transcription initiation are fairly well characterized; by contrast, the mechanisms by which insulator and silencer elements buffer or repress gene regulation, respectively, are less well understood[6]. Transcriptional enhancers are regulatory sequences that can be located upstream of, downstream of or within their target gene and can modulate expression independently of their orientation[7]. In vertebrates, enhancer sequences are thought to comprise densely clustered aggregations of transcription-factor-binding sites[8]. When appropriate occupancy of transcription-factor-binding sites is achieved, recruitment of transcriptional coactivators and chromatin-remodelling proteins occurs. The resultant protein aggregates are thought to facilitate DNA looping and ultimately promoter-mediated gene activation (see page 199). In-depth studies of individual genes such as *APOE* or *NKX2-5* (reviewed in ref. 9) have shown that many genes are regulated by complex arrays of enhancers, each driving distinct aspects of the messenger RNA expression pattern. These modular properties of mammalian enhancers are also supported by their additive regulatory activities in heterologous recombination experiments[10].

The purely genetic evidence from GWAS does not allow any direct inferences regarding the underlying molecular mechanisms, but a number of in-depth studies of individual loci (see below) suggest that variation in distant-acting enhancer sequences and the resultant changes in their activities can contribute to human disorders. Although we anticipate a variety of other non-coding functional categories such as negative gene regulators or non-coding RNAs to have a role in human disease, in this Review we focus on the role of enhancers and on strategies to define their location and function throughout the genome.

Enhancers in human disease

Beginning with the discovery that an inherited change in the β-globin gene alters one of the coded amino acids and thereby causes sickle-cell anaemia[11,12], thousands of mutations in the coding regions of genes have been identified to be responsible for monogenic disorders over the past half century. By contrast, the role of mutations not involving primary gene structural sequences has been minimally explored, largely owing to our inability to recognize relevant non-coding sequences, much less predict their function. The molecular genetic identification of individual enhancers involved in disease has been, in most cases, a painstaking and inefficient endeavour. Nevertheless, a number of successful studies have shown that distant-acting gene enhancers exist in the human genome and that variation in their sequences can contribute to disease. In this section, we discuss three examples in which enhancers were directly shown to play a role in human disease: thalassaemias resulting from deletions or rearrangements of β-globin gene (*HBB*) enhancers, preaxial polydactyly resulting from sonic hedgehog (*SHH*) limb-enhancer point mutations, and susceptibility to Hirschsprung's disease associated with a *RET* proto-oncogene enhancer variant.

The extensive studies of the human globin system and its role in haemoglobinopathies have historically served as a test bed for defining not only the role of coding sequences in disease[11,12] but also that of non-coding

[1]Genomics Division, MS 84–171, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [2]US Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA.

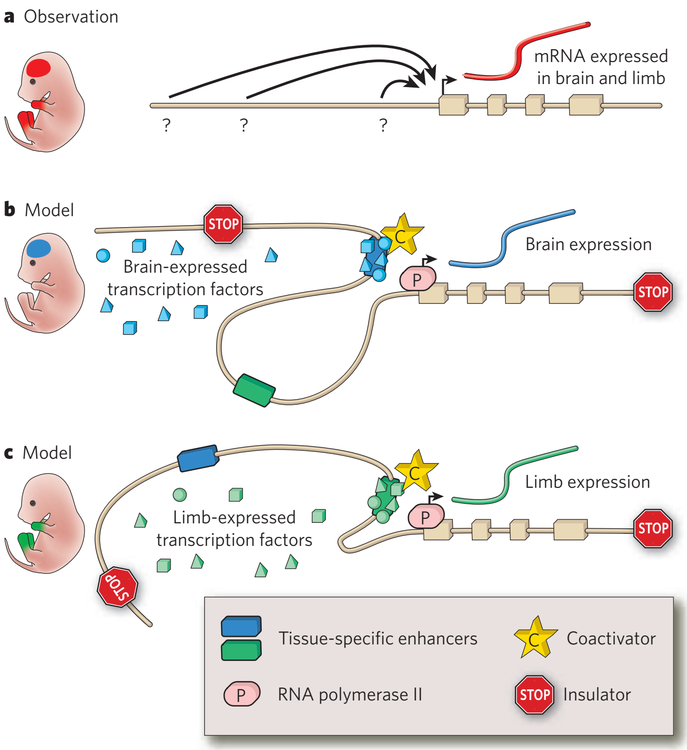

a Observation
mRNA expressed in brain and limb
?
?
?
b Model
STOP
Brain-expressed transcription factors
C
P
Brain expression
STOP
c Model
Limb-expressed transcription factors
C
P
Limb expression
STOP
STOP
Tissue-specific enhancers
C
Coactivator
P
RNA polymerase II
STOP
Insulator

Figure 1 | Overview of gene regulation by distant-acting enhancers. a, For many genes, the regulatory information embedded in the promoter is insufficient to drive the complex expression pattern observed at the messenger RNA level. For example, a gene could be expressed both in the brain and in the limbs during embryonic development (red), even if the promoter by itself is not active in either of these structures, suggesting that appropriate expression depends on additional sequences that are distant-acting and *cis*-regulatory. However, defining the genomic locations of such regulatory elements (question marks) and their activities in time and space (arrows) is a major challenge. **b**, **c**, 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 does the enhancer become active: it binds to transcriptional coactivators, relocates into physical proximity with 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 schematically shown in **b** and **c** for the example gene pictured in **a**, whose expression is controlled by two separate enhancers with brain-specific 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).

sequences. The α-thalassaemias and β-thalassaemias are haemoglobinopathies resulting from imbalances in the ratio of α-globin to β-globin chains in red blood cells. The molecular basis of these conditions was initially elucidated in cases in which inactivation or deletion of globin structural genes could be readily identified[13]. However, although gene deletion or sequence changes resulting in a truncated or non-functional gene product explained some thalassaemia cases, for a subset of patients intensive sequencing efforts failed to reveal abnormalities in globin protein-coding sequences. Through extensive long-range mapping and sequencing of DNA from individuals diagnosed with thalassaemia but lacking globin coding mutations, it was eventually discovered that many of these globin chain imbalances were due to deletion or chromosome rearrangements that resulted in the repositioning of distant-acting enhancers required for normal globin gene expression[14,15]. These early molecular genetic studies revealed a clear role for non-coding regulatory elements as a cause of human disorders through their impact on gene expression. Since then, many such examples of 'position effects', defined as changes in the expression of a gene when its location in a chromosome is changed, often by translocation, have been found[16].

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 non-coding mutation involves the limb-specific long-distance enhancer ZRS (also known as MFCS1) of *SHH* (Fig. 2). This enhancer is located at the extreme distance of approximately 1 megabase from *SHH*, within the intron of a neighbouring gene[17,18]. Of interest is that, initially, the gene in which the enhancer resides was thought to be relevant for limb development and was therefore named limb region 1 (*LMBR1*)[19]. 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[18,20]. Studies of the impact of the human ZRS sequence changes have been carried out in transgenic mice, in which the single-nucleotide changes result in ectopic anterior-limb expression during development, consistent with preaxial digit outgrowth[21]. Furthermore, sequence changes in the orthologous enhancers were found in mice, as well as in cats, with preaxial polydactyly[22,23], and targeted deletion of the enhancer in mice caused truncation of limbs[17]. These studies illustrate the importance of first experimentally identifying distant-acting enhancers in allowing subsequent human genetic studies to explore the potential role of disease-causing mutation in functional non-coding sequences.

Another example of enhancer variation contributing to human disease is provided by the discovery of a common non-coding variant linked to susceptibility to Hirschsprung's disease. Although multigenic, Hirschsprung's disease risk is strongly linked to coding mutations in the *RET* proto-oncogene[24,25]. However, family-based studies have also revealed evidence for Hirschsprung's disease linked to the *RET* locus in people lacking any accompanying functional *RET* coding mutations. Through the use of multispecies comparisons of orthologous genomic intervals that include and flank *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 more than a 20-fold increased risk for Hirschsprung's disease than rarer alleles in this element[26,27]. In transgenic mice, this enhancer was shown to be active in the nervous system and digestive tract during embryogenesis in a manner consistent with its putative role in Hirschsprung's disease[27]. It is interesting to note that although this enhancer variation is clearly important in disease risk, the variant alone is not sufficient to cause Hirschsprung's disease, highlighting the complex aetiology of this disorder.

As is evident from these labour-intensive gene-centric studies, enhancers can, in principle, have an important role in disease, but it remains unclear whether these are rare exceptions or whether variation in enhancers contributes to disease on a pervasive scale. Support for the latter comes from a rapidly growing number of examples in which non-coding SNPs linked to disease traits through GWAS were found to affect the expression levels of nearby genes[28], suggesting that variation in regulatory sequences may commonly contribute to a wide range of disorders. The results of the recent GWAS, 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 to defining the general characteristics of gene regulatory regions in specific human disorders, but they 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 made possible 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[29–32]. Thus, it was proposed that statistical measures of evolutionary sequence constraint would provide a way to identify potential enhancer sequences within the vast amount of non-coding sequence in the human genome. Support for this approach initially came from retrospective comparative genomic analyses of experimentally well-defined enhancers; these analyses revealed that enhancers 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 move away from blind studies of regions flanking genes of interest towards focusing specifically on non-coding sequences constrained across vertebrate species, culminating in whole-genome studies in which conservation level alone guided experimentation[32–34].

Initially, comparisons over extreme evolutionary distances, such as between humans and fish, were deemed most effective for this purpose[29,31]. Indeed, it was observed through large-scale transgenic mouse and fish studies that many of these non-coding sequences that had been conserved for hundreds of millions of years of evolution were enhancers that drove expression in highly specific anatomical structures during embryonic development. Likewise, so-called ultraconserved non-coding elements, which are blocks of 200 base pairs or more that are perfectly conserved between humans, mice and rats[35], 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[32]. This idea was further supported by the development of advanced statistical tools designed to quantify evolutionary constraint, from which it became evident that even comparisons between relatively closely related species can be effective predictors of enhancers[2,36,37]. A large-scale transgenic mouse study that included nearly all non-exonic ultraconserved elements in the human genome revealed that whereas many of them are developmental *in vivo* enhancers, other conserved non-coding sequences that are under similar evolutionary constraint, but are not perfectly conserved between humans and mice, are equally enriched in enhancers[33]. These results suggest that ultraconserved elements do not represent a functionally distinct subgroup of conserved non-coding sequences in terms of their enrichment in *in vivo* enhancers but rather that there is a much larger number of non-coding sequences that are under similar evolutionary constraint and are just as enriched in enhancers as are ultraconserved elements.

Independent of the specific algorithms and metrics that were used, most categories of conserved non-coding sequence were found not to be randomly distributed in the genome. Instead, they are located in a highly biased manner near genes active during development[2,33–35], consistent with the observation that a large proportion of these non-coding sequences give robust positive signals in various assays of being tissue-specific *in vivo* enhancers active during development.

Comparative approaches are an effective high-throughput genomic strategy for identifying non-coding sequences that are highly likely to be enhancers, but they have several limitations. First, although conservation is indicative of function, it is not necessarily indicative of enhancer activity, because many other types of non-coding functional element that may have similar conservation signatures are known to exist. Second, even when conservation of non-coding DNA results from 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 data sets.

A perplexing study questioning the importance of extremely conserved enhancers found the lack of an apparent phenotype upon targeted deletion of four independent ultraconserved elements in mice[38]. General expectations were that non-coding sequences that have been perfectly conserved in mammals for tens of millions of years must be essential and that their deletion should result in severe phenotypes, comparable to those observed upon deletion of the *Shh* limb enhancer and other less well-conserved enhancers[9,17]. However, mice with deletions of such ultraconserved enhancers were viable, fertile and showed no overt phenotype[38]. Interpretations of this lack of obvious effect are similar to those of the absence of phenotypes upon deletion of highly conserved protein-coding genes: minor phenotypes may have escaped detection in the assays used; there may have been functional redundancy with other genes or enhancers; or there may have been reductions in fitness that only become apparent over multiple generations or are not easily detected in a controlled laboratory environment. This study highlighted that although extreme non-coding 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 strategy complementary 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

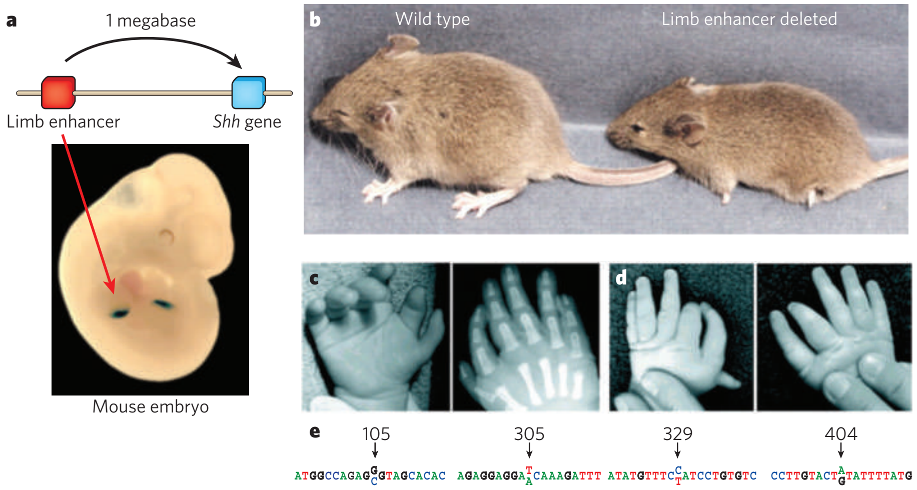

a
1 megabase
Limb enhancer
Shh gene
Mouse embryo
b
Wild type
Limb enhancer deleted
c
d
e
105
305
329
404
ATGGCCAGAG$\overset{G}{\underset{C}{}}$GTAGCACAC
AGAGGAGGA$\overset{T}{\underset{A}{}}$CAAAGATTT
ATATGTTTC$\overset{C}{\underset{T}{}}$ATCCTGTGTC
CCTTGTACT$\overset{A}{\underset{G}{}}$TATTTTATG

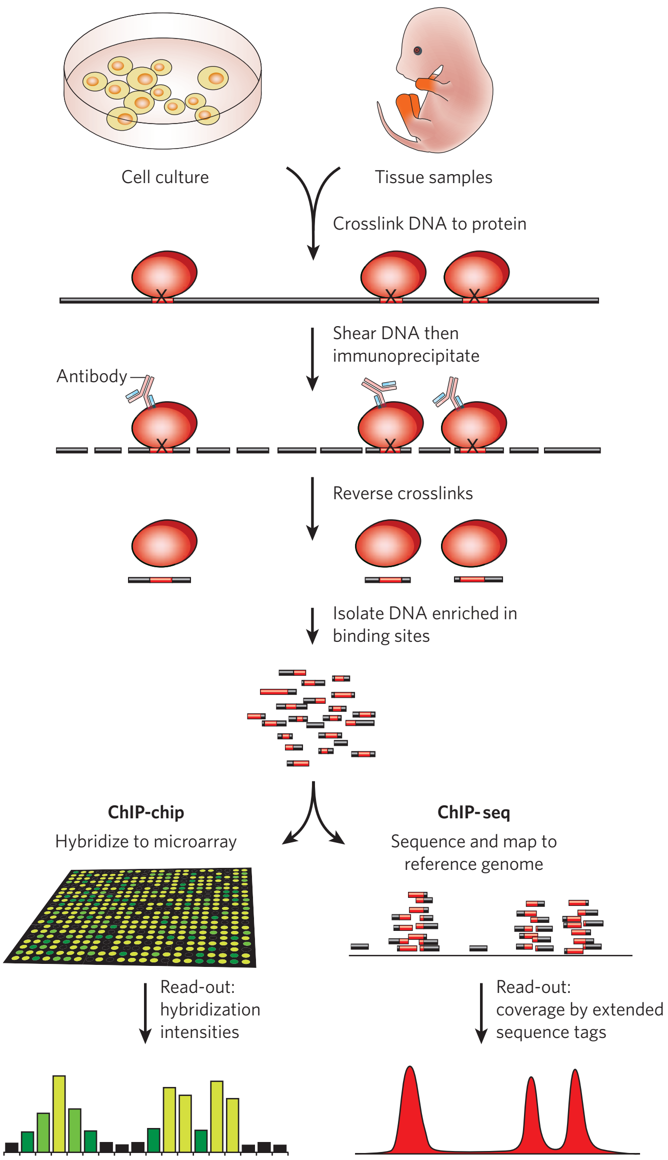
Figure 2 | Consequences of deletion and mutation of the limb enhancer of sonic hedgehog. a, The limb enhancer of *Shh* is located approximately 1 megabase away from its target promoter in the intron of a neighbouring gene (*Lmbr1*; exons not shown). In transgenic mouse reporter assays, this non-coding sequence targets gene expression to a posterior region of the developing limb bud (red arrow). (Image reproduced, with permission, from ref. 18.) **b**, Mice with a targeted deletion of this enhancer have severely truncated limbs, which strikingly demonstrates its functional importance in development. (Reproduced, with permission, from ref. 17.) **c**–**e**, Point mutations in the orthologous human enhancer sequence result in preaxial polydactyly, emphasizing the potential significance of variation in non-coding functional sequences in both rare and common human disorders: **c** and **d** show the hands of two patients with point mutations in the *SHH* limb enhancer; **e** shows point mutations associated with preaxial polydactyly identified in four unrelated families. (Panels **c** and **d** reproduced, with permission, from ref. 18; panel **e** modified, with permission, from ref. 18.)

Box 1 | Mapping of regulatory elements using ChIP-chip and ChIP-seq

Formaldehyde crosslinking of DNA to proteins that bind to it directly or as part of larger complexes[71], combined with subsequent immunoprecipitation targeting specific DNA-associated proteins (ChIP[72]), was widely used in the pre-genomic era to study protein–DNA interactions directly in cultured cells or in tissue samples. The top portion of the figure shows a schematic overview of the individual steps involved. They include the molecular fixation of non-covalent protein–DNA interactions, shearing of the crosslinked chromatin, immunoprecipitation with an antibody binding the protein of interest and reversal of crosslinks. In many cases, antibodies that bind to covalently modified proteins are used, for example those that recognize methyl groups at defined amino-acid residues of histones. In the conventional ChIP approach, enrichment of the associated DNA fragments relative to non-immunoprecipitated ('input') DNA is quantified for individual proposed binding locations (not shown). This need for quantification at every site of interest initially thwarted the application of ChIP on a genomic scale.

The introduction of DNA microarrays allowed the hybridization-based interrogation of large numbers of potential binding sites in parallel (ChIP-on-chip, or ChIP-chip), thus making it possible to screen entire compact model-organism genomes[73,74] or large vertebrate genome intervals[75] in a single experiment (see figure, bottom left). ChIP-chip was used on a massive scale in the Encyclopedia of DNA Elements (ENCODE) pilot project, in which dozens of proteins and protein modifications were initially mapped in a representative 1% portion of the human genome[39].

Recently, chromatin immunoprecipitation coupled to massively parallel sequencing (ChIP-seq) has become increasingly used as an alternative to ChIP-chip[44-47]. The ChIP-seq method is very similar to the experimental set-up of ChIP-chip, except that, in the final step, next-generation sequencing techniques are used to determine the sequence of immunoprecipitated DNA fragments, which are then computationally mapped to the reference genome (see figure, bottom right). Improved sequencing technologies offer the possibility to obtain millions of mappable reads in a single ChIP-seq experiment at moderate cost. The results from ChIP-seq are based on statistical analysis of read counts, which overcomes many of the challenges associated with the quantification and normalization of hybridization signals, and an increasing number of advanced computational ChIP-seq analysis tools are becoming available[76]. ChIP-seq analysis covers by default the entire mappable portion of the reference genome without the need to restrict the analysis to its subregions.


Cell culture
Tissue samples
Crosslink DNA to protein
Shear DNA then immunoprecipitate
Antibody
Reverse crosslinks
Isolate DNA enriched in binding sites
ChIP-chip
Hybridize to microarray
ChIP-seq
Sequence and map to reference genome
Read-out: hybridization intensities
Read-out: coverage by extended sequence tags

regulatory regions. These genomic approaches have become possible as a result of an improved understanding of the proteins and epigenetic marks found at particular categories of regulatory element, together with concurrently developed technologies that allow traditional chromatin immunoprecipitation (ChIP) techniques to be applied on the scale of whole vertebrate genomes. The initial in-depth studies of 1% of the genome in the Encyclopedia of DNA Elements (ENCODE) pilot project[39] were largely based on data sets generated by the ChIP-chip technique (Box 1) and revealed the molecular properties of a variety of regulatory elements.

With respect to enhancer identification, a particularly relevant insight was the identification of specific histone 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 at this position (H3K4me1)[40]. Mapping these marks in the ENCODE regions and, more recently, throughout the entire genome[41] 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 (also known as MED1), raising the possibility that such coactivators might be useful general markers for mapping enhancers. Although it was initially not clear to what extent the presence of transcriptional coactivators such as p300 is indicative of active rather than inactive enhancers, comparison of DNase I hypersensitivity (a marker of open chromatin structure) in several cell lines throughout the ENCODE regions revealed that the location of cell-line-specific distal DNase-I-hypersensitivity 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 the mapping of DNA elements with cell-specific and tissue-specific enhancer activities[42].

Owing to the development of the ChIP-seq technique (Box 1), 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 humans and mice have become available[43–55]. These data sets allowed the identification of not only the H3K4me1 and H3K4me3 signatures discussed earlier but also additional chromatin marks present at predicted or validated enhancers, and provided a refined view of their correlation to enhancer activities[44,51,55]. However, with very few exceptions (see, for example, refs 50 and 54) genome-wide mapping of these and other regulation-associated chromatin marks (Table 1) was done in immortalized cell lines, cultured stem cells or primary cell cultures. Thus, the maps of potentially enhancer-associated marks produced by these studies provided limited insight into their *in vivo* distribution during embryonic development and in adult organs, most probably 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[56]. 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 non-coding regions distal from known promoters. Transgenic mouse experiments with almost 100 of these sequences revealed that they are developmental enhancers in almost all cases. 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. The study also showed global enrichment in tissue-specific p300 peaks near genes that are expressed in the same tissue, again consistent with the proposed function of these genomic regions as active transcriptional enhancers.

These experimentally predicted genome-wide sets of *in vivo* enhancers also made it possible to address the controversial issue of the extent to which evolutionary conservation is a hallmark of *in vivo* enhancers[57]. Several studies have shown that highly conserved non-coding elements are enriched in developmental *in vivo* enhancers[32–34]. However, some observations have challenged such a generalized correlation between sequence conservation and enhancer activity: experimental analysis of individual loci suggested that a large proportion of enhancers cannot be detected by comparative genomics[58]; the molecular marks of a surprisingly large proportion of sequences in the ENCODE regions suggested that regulatory functions are not, or are only weakly, conserved[39]; and histone methylation present at orthologous loci in humans and mice did not correlate with overall increased levels of sequence conservation[59]. 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[56]. There may be variation in the degree of evolutionary constraint of enhancers that are active in different types of cell or developing tissue, but these data suggest that developmental enhancers that can be identified through p300 binding are commonly evolutionarily constrained.

Although preliminary, 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 increase and the cost reductions of next-generation sequencing technologies offer an increasingly powerful genome-wide means of identifying specific DNA–protein interactions. 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 states, as well as diseased states, providing multilayered *in vivo* annotations of the non-coding portion of our genome. It is important to 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 study conclusively the mechanisms by which sequence variation in distant-acting enhancers contributes to disease.

Defining the targets

The methods described here have considerably improved our ability to identify enhancers and their associated activity patterns on a genomic scale, but a remaining important challenge is to determine the relationships between enhancers and genes. Comparing ChIP-chip or ChIP-seq data with transcriptome data from microarrays or RNA-seq[60] can provide highly suggestive clues to the identity of the target gene of a given enhancer in a given tissue, but such comparisons do not provide the direct evidence for enhancer–promoter interactions that would be desirable in mapping 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, but it only became possible to study such interactions systematically through the introduction of the chromosome conformation capture (3C) assay and its derivative technologies[61]. Similar to ChIP, the 3C approach relies on formaldehyde crosslinking to capture DNA–DNA interactions directly in intact cells or cell nuclei. Previously suggested pairs of interacting sites are subsequently tested and validated one by one through the quantification of crosslinking events. In one of many examples demonstrating the utility of 3C in the analysis of distant-acting vertebrate enhancers, this technique was recently used[62] to study chromatin interactions at the *Shh* locus, whose role in limb development was discussed in detail earlier. Using the 3C technique, it was demonstrated 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 non-coding sequences and actively transcribed genes; this was first demonstrated in the mouse β-globin gene locus[63].

This work and other gene-centric studies (for more examples, see refs 64 and 65) were critical in shaping our understanding of enhancer–promoter interactions. However, they have the fundamental limitation that only one or very few previously proposed interactions between specific loci can be assayed per experiment. This limitation was partly overcome through the use of microarrays to analyse entire 3C libraries (called chromosome conformation capture-on-chip[66] and circular chromosome conformation capture[67], both known as 4C). By applying this approach to fetal liver and brain, it was demonstrated that the β-globin gene locus control region (LCR) makes reproducible tissue-specific contacts with other loci predominantly located on the same chromosome but in some cases dozens of megabases away from the LCR[66]. Of possible relevance to the adoption of this approach for enhancer discovery is that reproducible interactions with other chromosome regions were also observed in the brain, where the LCR is thought to be inactive.

The 4C approaches are 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 partly alleviated by the chromosome conformation capture carbon copy (5C) method[68], in which a complex 3C library generated through multiplexed PCR is analysed by large-scale sequencing to generate a comprehensive 'many-to-many' interaction map of DNA–DNA interactions. However, owing to the need for specific

Table 1 | Selected major categories of non-coding functional element

Category	Function	Selected associated chromatin marks*
Promoter	Region that is located immediately upstream of a protein-coding gene, and binds to RNA polymerase II; where transcription is initiated	RNA polymerase II[44], H3K4me3 (ref. 40) (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 from target promoters and are orientation independent	p300 (refs 40, 56), H3K4me1 (ref. 40)
Insulator	Separates active from inactive chromatin domains and interferes with enhancer activity when placed between an enhancer and promoter	CTCF[44,53]
Repressor/ silencer	Negative regulators of gene expression	REST[45], SUZ12 (refs 69, 70)

*Many additional chromatin marks were found to correlate with one or several of these categories of regulatory element. Detailed descriptions of these markers and their respective binding characteristics at different types of regulatory sequence element can be found in refs 40, 41, 44, 51 and 55.

primers for each possible interacting fragment and the sequencing depth required for analysis of the resultant libraries, the 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, antibody-based methods might be used to restrict the analysis space in which DNA–DNA interactions are studied to a size that can be affordably analysed using currently available sequencing technologies. One possibility is to couple a chromatin-interaction paired-end tag (ChIA–PET) sequencing strategy to a ChIP step that enriches for chromatin fragments bound to a specific transcription factor or other chromatin mark of interest[64]. Although the technical feasibility of this approach remains to be demonstrated, it has remarkable potential for enhancer discovery. This is because its application to general enhancer-associated marks such as p300 or histone methylation[40,56] might identify, in a single step, enhancers active in a tissue of interest, as well as their respective target genes.

Perspective

Genetic and medical resequencing studies have been advanced by knowledge about the structure of protein-coding genes and a detailed understanding of the relationship between mRNA sequences and the primary structures of the proteins they encode. Through such studies, disease links have been established for a sizeable proportion of the ~20,000 protein-coding genes in the human genome. By contrast, a very limited number of changes in gene regulatory sequences have so far been linked to human disease. Consequently, an important motivation for functionally annotating the non-coding portion of the human genome and the *cis*-regulatory elements that it contains is to assess the relationship between variations in non-coding sequences and human disease. In the absence of genome-wide catalogues of functionally annotated regulatory elements, how these elements impact on human biology, as well as disease, will remain an untested hypothesis.

Despite advances in relevant technologies, functionally characterizing the distant-acting-enhancer architecture of the human genome in its entirety will be an enormous undertaking, owing to the great number of data points needed, which include dozens of tissues and cell types, as well as developmental states and possibly disease states.

A further challenge will be to link distant-acting enhancers to the genes they regulate. Linking enhancers to their cognate gene will allow the further assignment of these functional sequences to their basic 'gene' unit of heredity, for collective resequencing analysis.

Although we have focused on distant-acting enhancers here, there are other categories of functional element in the non-coding portion of the genome (for example insulators, negative regulators, promoters and non-coding RNAs), and they will also be crucial targets for large-scale identification and characterization. It is expected that technologies similar to those described here for enhancers will make it possible to explore their roles in human biology and disease. ■

1. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
2. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
3. Helgadottir, A. *et al.* A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316,** 1491–1493 (2007).
4. McPherson, R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316,** 1488–1491 (2007).
5. Hindorff, L. A., Junkins, H. A., Mehta, J. P. & Manolio, T. A. A catalog of published genome-wide association studies. *OPG: Catalog Published Genome-Wide Assoc. Studies* <http://www.genome.gov/gwastudies> (2009).
6. 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).
   This paper is a comprehensive overview of functional classes of gene regulatory sequence, including many disease-relevant examples identified through gene-centric studies.
7. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27,** 299–308 (1981).
8. Panne, D. The enhanceosome. *Curr. Opin. Struct. Biol.* **18,** 236–242 (2008).
9. Visel, A., Bristow, J. & Pennacchio, L. A. Enhancer identification through comparative genomics. *Semin. Cell Dev. Biol.* **18,** 140–152 (2007).
10. Visel, A. *et al.* Functional autonomy of distant-acting human enhancers. *Genomics* **93,** 509–513 (2009).
11. Ingram, V. M. Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature* **180,** 326–328 (1957).
12. Pauling, L. *et al.* Sickle cell anemia, a molecular disease. *Science* **110,** 543–548 (1949).
13. Kan, Y. W. *et al.* Deletion of α-globin genes in haemoglobin-H disease demonstrates multiple α-globin structural loci. *Nature* **255,** 255–256 (1975).
14. Kioussis, D., Vanin, E., deLange, T., Flavell, R. A. & Grosveld, F. G. β-Globin gene inactivation by DNA translocation in γβ-thalassaemia. *Nature* **306,** 662–666 (1983).
15. Semenza, G. L. *et al.* The silent carrier allele: β thalassemia without a mutation in the β-globin gene or its immediate flanking regions. *Cell* **39,** 123–128 (1984).
16. Kleinjan, D. A. & Lettice, L. A. Long-range gene control and genetic disease. *Adv. Genet.* **61,** 339–388 (2008).
17. 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,** 797–803 (2005).
    This paper shows that deletion of the distant-acting limb enhancer of the *Shh* gene in mice causes severe limb truncation, providing a model example of the requirement for enhancers in mammalian development.
18. 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,** 1725–1735 (2003).
19. 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,** 19–27 (2000).
20. 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,** 2417–2423 (2008).
21. 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,** 207–214 (2007).
22. 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,** 978–985 (2008).
23. Lettice, L. A. *et al.* Disruption of a long-range *cis*-acting regulator for *Shh* causes preaxial polydactyly. *Proc. Natl Acad. Sci. USA* **99,** 7548–7553 (2002).
24. 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. USA* **97,** 268–273 (2000).
25. Gabriel, S. B. *et al.* Segregation at three loci explains familial and population risk in Hirschsprung disease. *Nature Genet.* **31,** 89–93 (2002).
26. Emison, E. S. *et al.* A common sex-dependent mutation in a *RET* enhancer underlies Hirschsprung disease risk. *Nature* **434,** 857–863 (2005).
27. 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,** 3837–3845 (2005).
28. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nature Rev. Genet.* **10,** 184–194 (2009).
29. Aparicio, S. *et al.* Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proc. Natl Acad. Sci. USA* **92,** 1684–1688 (1995).
30. Loots, G. G. *et al.* Identification of a coordinate regulator of interleukins 4, 13 and 5 by cross-species sequence comparisons. *Science* **288,** 136–140 (2000).
31. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302,** 413 (2003).
32. Pennacchio, L. A. *et al. In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444,** 499–502 (2006).
33. Visel, A. *et al.* Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nature Genet.* **40,** 158–160 (2008).
34. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3,** e7 (2005).
35. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304,** 1321–1325 (2004).
36. Prabhakar, S. *et al.* Close sequence comparisons are sufficient to identify human *cis*-regulatory elements. *Genome Res.* **16,** 855–863 (2006).
37. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–913 (2005).
38. Ahituv, N. *et al.* Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* **5,** e234 (2007).
    This paper shows that deletion of several ultraconserved non-coding sequences in mice may not result in obvious phenotypes, demonstrating that even extreme evolutionary constraint does not necessarily indicate that a non-coding sequence is required for viability.
39. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
40. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39,** 311–318 (2007).
    This paper identifies a histone H3K4 differential methylation signature that distinguishes promoters from enhancers, providing a chromatin-based tool for genome-wide enhancer prediction.
41. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 (2009).
42. Xi, H. *et al.* Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet.* **3,** e136 (2007).
43. Wei, C. L. *et al.* A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124,** 207–219 (2006).
    This paper describes mapping of protein–DNA interactions by ChIP coupled with conventional capillary-based sequencing of concatenated paired-end tags (ChIP-PET), a conceptual predecessor of the ChIP-seq approach.
44. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).
45. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316,** 1497–1502 (2007).

46. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4,** 651–657 (2007).
This paper is one of several independently published early ChIP-seq studies validating the method for genome-wide mapping of transcription-factor-binding sites.
47. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
This paper is one of several independently published early ChIP-seq studies providing some of the first genome-wide data sets of several histone modifications in different mouse cell types and examining their correlation with functional genome features.
48. 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,** 286–298 (2007).
49. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133,** 1106–1117 (2008).
50. 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,** 4549–4564 (2008).
51. Robertson, A. G. *et al.* Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res.* **18,** 1906–1917 (2008).
52. Ku, M. *et al.* Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.* **4,** e1000242 (2008).
53. 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,** 24–32 (2009).
54. Gao, N. *et al.* Dynamic regulation of *Pdx1* enhancers by Foxa1 and Foxa2 is essential for pancreas development. *Genes Dev.* **22,** 3435–3448 (2008).
55. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genet.* **40,** 897–903 (2008).
56. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457,** 854–858 (2009).
57. Cooper, G. M. & Brown, C. D. Qualifying the relationship between sequence conservation and molecular function. *Genome Res.* **18,** 201–205 (2008).
58. McGaughey, D. M. *et al.* Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at *phox2b*. *Genome Res.* **18,** 252–260 (2008).
59. Bernstein, B. E. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120,** 169–181 (2005).
60. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10,** 57–63 (2009).
61. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295,** 1306–1311 (2002).
62. Amano, T. *et al.* Chromosomal dynamics at the *Shh* locus: limb bud-specific differential regulation of competence and active transcription. *Dev. Cell* **16,** 47–57 (2009).
63. Carter, D., Chakalova, L., Osborne, C. S., Dai, Y. F. & Fraser, P. Long-range chromatin regulatory interactions *in vivo*. *Nature Genet.* **32,** 623–626 (2002).
64. 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,** 521–532 (2009).
65. Miele, A. & Dekker, J. Long-range chromosomal interactions and gene regulation. *Mol. Biosyst.* **4,** 1046–1057 (2008).
66. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genet.* **38,** 1348–1354 (2006).
67. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genet.* **38,** 1341–1347 (2006).
68. Dostie, J. *et al.* Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16,** 1299–1309 (2006).
69. Lee, T. I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125,** 301–313 (2006).
70. Squazzo, S. L. *et al.* Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res.* **16,** 890–900 (2006).
71. Van Lente, F., Jackson, J. F. & Weintraub, H. Identification of specific crosslinked histones after treatment of chromatin with formaldehyde. *Cell* **5,** 45–50 (1975).
72. 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,** 937–947 (1988).
73. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290,** 2306–2309 (2000).
74. Iyer, V. R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409,** 533–538 (2001).
75. Horak, C. E. *et al.* GATA-1 binding sites mapped in the β-globin locus by using mammalian chIp-chip analysis. *Proc. Natl Acad. Sci. USA* **99,** 2924–2929 (2002).
76. Barski, A. & Zhao, K. Genomic location analysis by ChIP-seq. *J. Cell. Biochem.* **107,** 11–18 (2009).

Acknowledgements We thank M. Blow, S. Deutsch and A. Sczyrba for help with computational analysis of GWAS data and C. Attanasio for comments. L.A.P. and E.M.R. were supported by the Berkeley Program for Genomic Applications (funded by the US National Heart, Lung, and Blood Institute), and the Director, Office of Science, Office of Basic Energy Sciences, US Department of Energy, under contract number DE-AC02-05CH11231. L.A.P. was also supported by the US National Human Genome Research Institute.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to L.A.P. (lapennacchio@lbl.gov).