Published in final edited form as:

Nat Rev Genet. 2013 April ; 14(4): 288–295. doi:10.1038/nrg3458.

Enhancers: five essential questions

Len A. Pennacchio,

Genomics Division, One Cyclotron Road, MS 84–171, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

Wendy Bickmore,

MRC Human Genetics Unit, IGMM, University of Edinburgh, Crewe Road, Edinburgh EH4 2XU, UK

Ann Dean,

Laboratory of Cellular and Developmental Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, 50 South Drive, MSC 8028, Bethesda, Maryland 20892, USA

Marcelo A. Nobrega, and

Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA

Gill Bejerano

Developmental Biology and Computer Science, Stanford University, Beckman Center B-300, 279 Campus Drive West (MC 5329), Stanford, California 94305–5329, USA

Len A. Pennacchio: lapennacchio@lbl.gov; Wendy Bickmore: wendy.bickmore@igmm.ed.ac.uk; Ann Dean: anndean@helix.nih.gov; Gill Bejerano: bejerano@stanford.edu

Abstract

It is estimated that the human genome contains hundreds of thousands of enhancers, so understanding these gene-regulatory elements is a crucial goal. Several fundamental questions need to be addressed about enhancers, such as how do we identify them all, how do they work, and how do they contribute to disease and evolution? Five prominent researchers in this field look at how much we know already and what needs to be done to answer these questions.

Q *What are the challenges in identifying all enhancers and their functions?*

Correspondence: nobrega@uchicago.edu.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Wendy Bickmore's homepage: http://www.hgu.mrc.ac.uk/people/w.bickmore.html
Ann Dean's homepage: http://www2.niddk.nih.gov/NIDDKLabs/IntramuralFaculty/DeanAnn.htm
Marcelo A. Nobrega's homepage: http://genes.uchicago.edu/contents/faculty/nobrega-marcelo.html
Gill Bejerano's homepage: http://bejerano.stanford.edu
A database of *in vivo* enhancer studies: http://enhancer.lbl.gov
***Nature Reviews Genetics* Series on Regulatory elements:** http://www.nature.com/nrg/series/regulatoryelements/index.html
ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Len A. Pennacchio. Enhancers are classically defined as *cis*-acting DNA sequences that can increase the transcription of genes. They generally function independently of orientation and at various distances from their target promoter (or promoters). Historically, the identification of enhancers has proved challenging for several reasons[1]. First, enhancers are scattered across the 98% of the human genome that does not encode proteins, resulting in a large search space (billions of base pairs of DNA). Second, while it is known that they regulate genes in *cis*, their location relative to their target gene (or genes) is highly variable: namely, enhancers can be found upstream or downstream of genes but also within introns. Furthermore, they do not necessarily act on the respective closest promoter but can bypass neighbouring genes to regulate genes located more distantly along a chromosome. And in some cases, individual enhancers have been found to regulate multiple genes[2], adding further complexity to their functional annotation. Third, in contrast to the well-defined sequence code of protein-coding genes, the general sequence code of enhancers, if one exists at all, is poorly understood. Thus, enhancers cannot be identified computationally from DNA sequence alone with high confidence. Finally, the activity of enhancers can be restricted to a particular tissue or cell type, a time point in life, or to specific physiological, pathological or environmental conditions. While this dynamic nature of enhancers enables their genomic function to determine precisely when, where and at what level each of our genes is expressed, it further complicates the discovery and functional annotation of enhancers in the genome.

Despite these challenges, traction has been made in the past decade for identifying enhancers on a genome scale. Initially, this was facilitated by comparative genomics, in which non-coding sequences that are highly conserved between different vertebrate and mammalian species were found to be enriched for enhancers, especially those that are active in early development[3,4]. Remarkably, the systematic testing of hundreds of highly conserved human non-coding DNA sequences in transgenic mouse reporter assays revealed that about half were enhancers[4,5]. This enrichment in enhancers is surprising, since these studies assayed only a single time point of mouse development (namely, embryonic day 11.5), and in principle numerous other functions exist in non-coding DNA that could cause sequences to be conserved. These findings suggested that enhancers are a major category and are possibly even the predominant type of functional element in the non-coding portion of the genome.

Even with the success of comparative genomics in identifying enhancers, conservation alone has limitations. For instance, there are hundreds of highly conserved genomic sequences for which no enhancer function could be demonstrated in transgenic assays. This might be due to limitations of currently available assays to selected time points of development, but it is also possible that these sequences are conserved because of important functions other than being enhancers. Furthermore, the fact that a sequence is conserved and therefore might be an enhancer does not provide any clues as to when and where such function might occur within humans. As an additional challenge, recent studies support that a substantial fraction of enhancers displays modest or no conservation across species, thereby further limiting this evolution-based approach[6,7].

Some of these challenges can be addressed using a more recent enhancer identification method that exploits advances in high-throughput sequencing of histone modifications and other epigenomic marks directly from cell lines or primary tissues. This so-called 'ChIP–seq' approach (for chromatin immunoprecipitation followed by high-throughput sequencing) proves to be powerful as it is independent of DNA conservation and defines enhancer catalogues directly from the cells or tissues of study. Marks commonly used for identifying putative enhancers include p300 (ref. 8), histone H3 acetylated at lysine 27 (H3K27ac)[9] and H3 monomethylated at K4 (H3K4me1)[10]. The mapping of DNase I hypersensitive sites represents another useful approach[11]. All of these marks have proved to be useful in various mammalian species, with antibodies to histone marks as well as DNase I hypersensitivity having broad application across many forms of eukaryotic life. These molecular tools were largely shown to be useful for enhancer identification through their experimental validation using enhancer reporter vectors either *in vitro*[10] or *in vivo*[8].

A remarkable finding from surveys of tissues and diverse cell lines is the growing evidence for the sheer number of enhancers in our genome. It is estimated that hundreds of thousands of enhancers exist in the human genome[12,13,14], vastly outnumbering our ~20,000 protein-encoding genes. This observation continues to point to the importance of regulating gene expression as a primary level of controlling genome and ultimately organism function.

A second finding from these epigenomic approaches is how poorly conserved enhancers can be in a given tissue. This includes examples of enhancers identified in liver tissue across a panel of vertebrates[6], as well as enhancers uncovered in heart tissues from both mice[7] and humans[15]. These early findings highlight the importance of studying certain tissues directly from the species under investigation (that is, humans) versus attempting to use standard animal models to derive their identity (that is, mice). It is anticipated that large studies aimed at enhancer identification, such as ENCODE[14], will continue to transition their focus from animal models and human cell lines to primary human tissues on the basis of these findings.

Despite the great value of this new generation of experimental tools for uncovering enhancers on a genome scale, there are also some limitations. For example, there is currently no single 'enhancer mark' that can be used to identify all genomic regions that are enhancers and that predicts with certainty whether a given enhancer is active or inactive in a given cell type or tissue. All enhancer prediction methods described to date, whether conservation- or epigenomics-based, are less than perfect; that is, comparison with experimental validation series shows that some enhancer regions are missed (false negatives), and other sequences predicted to be active enhancers cannot be validated by complementary methods (false positives). Consequently, in using such information for gene-centric follow-up, care must be taken to confirm that these predictions are valid prior to embarking on larger investigations. Notwithstanding the progress that has been made, complete annotation of all enhancers in the genome by epigenomic approaches remains a daunting task owing to the nearly endless number of cell types and conditions that one would need to explore. Further advances to fill this void are needed in higher-throughput and lower-cost enhancer identification strategies, as are advances in the ability to isolate and to work with smaller tissue input amounts (including single cells) as well as the parallel development of more effective computational

predictions of enhancers. Indeed, much work remains to identify enhancers globally and ultimately to connect their function to human biology and disease.

Q *How do enhancers interact with their targets in the complex three-dimensional environment of the genome?*

Wendy Bickmore. I think that we should break this question down into two parts. The first is 'do enhancers physically interact with their targets?' And, if so, then we can ask 'how?'

The idea that distant regulatory elements (enhancers) may exert their function by DNA looping originated from studies of bacterial regulators such as the *Escherichia coli lac* operator. While these elements work over relatively short (<100 bp) segments of DNA and on a non-nucleosomal template, this concept has been extrapolated to mammalian enhancers, which can be located as much as a million base pairs away from their target genes and function on a complex chromatin template.

Evidence for the formation of loops between long-range enhancers and promoters comes mainly from two strands of evidence. The first is the cross-linking, by formaldehyde, and subsequent ligation together of enhancer and promoter DNA sequences as detected in chromosome conformation capture (3C)-type methods. The second is the visualization, by fluorescence *in situ* hybridization (FISH), of the spatial proximity of enhancer and promoter regions in the cell nucleus. In some cases, both of these assays do indeed support looping mechanisms that bring distant *cis*-regulatory elements into very close (<200 nm) proximity of their target genes in a tissue-restricted manner.

However, in some other cases in which bona fide regulatory elements can be captured by cross-linking to the appropriate gene promoter, visual assays do not detect a significant frequency of spatial co-localization between the enhancer and the promoter[16]. This may be because the chromatin loops are too transient to detect by FISH or because the 3C ligation products are established through indirect cross-linking of enhancers and promoters to relatively large (300–400 nm) nuclear substructures or supramolecular complexes. In the latter case, I would say that there is not a DNA loop as such being formed between the enhancer and the promoter.

In cases where looping does occur, how do the loops form? Generally, the assumption is that enhancer–promoter looping serves to deliver factors (for example, RNA polymerase, transactivators and transcription factors) to the promoter in the right tissue and at the right time. Since chromatin is a very large flexible polymer, the default conformation of which is not a series of structured loops, there must be specific mechanisms for stable loop formation. The directed formation of large loops by active chromatin bending would require considerable energy input, and we do not know of active mechanisms operating in interphase that can do this over such large distances. However, chromatin continuously undergoes movement by constrained diffusion[17]. The radius of this constraint is sufficiently large that any two sequences within approximately 1 Mb of each other can randomly encounter each other in the nucleus. If there are protein complexes bound at the promoter and enhancer that have affinity for one another, a chromatin loop may then be stabilized through this passive mechanism. Proteins that might be able to do this include those with

dimerization or oligomerization domains and that are present at both the promoter and the enhancer. The most striking experimental demonstration of this, and of the ability of loops to activate transcription, comes from the tethering of LDB1 at the β-globin locus in erythroid cells[18].

The formation of a loop, which juxtaposes sequences associated with multiple transcription, chromatin-modifying and chromatin-remodelling factors, will increase the local concentration of these factors and so promote the formation of further protein– protein and protein–DNA complexes. Indeed, increased local protein concentration has been shown to be a key mechanism through which looping affects repression by the *lac* repressor[19].

So what about situations where DNA loops between enhancers and gene targets cannot be directly visualized in the nucleus? In some of these cases, the intervening chromatin seems to be in a compact state so that enhancer and promoter are still relatively close to each other (200–400 nm)[16]. The high concentrations of transcription factors and protein complexes nucleated by enhancer binding could then simply diffuse through this restricted nuclear volume to find and activate transcription from the target promoter. Diffusion might also be facilitated by nonspecific binding to the intervening chromatin, and indeed this type of scanning has been observed for the *lac* repressor *in vivo*[20]. Examples of proteins scanning the chromatin between enhancers and promoters have also been reported in eukaryotes[21]. Akin to the scanning model, the linking model proposes that chromatin complexes assembled at enhancers actively reorganize the chromatin between enhancers and promoters and is supported by evidence for propagation of histone modifications across the intervening chromatin[21] and by the activity of enhancer-blocking sequences[22].

It is harder to imagine the scanning and linking mechanisms operating at enhancers located hundreds of thousands of base pairs away from their target promoter, often with intervening genes that do not respond to the enhancer. Nor is there any reason to think that all enhancers function through the same mechanism. Indeed, components of both one-dimensional and three-dimensional diffusion have been observed for the *lac* repressor in living *E. coli* cells[20].

Q *How do enhancers bring about gene expression?*

Ann Dean. Enhancers are DNA-regulatory elements that activate transcription of a gene or genes to higher levels than would be the case in their absence. These elements function at a distance by forming chromatin loops to bring the enhancer and target gene into proximity[23]. It is thought that lineage-specific DNA-binding transcription factors bound at promoters and enhancers either interact with each other or recruit ‘looping’ factors that mediate the long-range contacts that are detected by chromosome conformation capture (3C) or related assays. Recent data also suggest that insulator-binding proteins CTCF and cohesin may facilitate enhancer–promoter interactions.

How do enhancers affect transcription? Genome profiling has revealed that general transcription factors (GTFs) and RNA polymerase II (Pol II) are recruited to enhancers[24]. Thus, it appears that enhancers serve as centres for the assembly of the pre-initiation complex (PIC). Loop formation might increase the local concentration of transcription machinery components in the vicinity of the target gene, or the enhancer might serve to

'deliver' the PIC to a promoter. Enhancers might be important for nuclear relocation of the enhancer–promoter pair to a neighbourhood that is favourable for transcription. There is evidence for each of these models, but important questions remain about the mechanistic details and how the models might relate to each other.

During PIC formation, the Mediator co-activator complex bridges upstream activators and Pol II. Can Mediator bridge to activators bound at enhancers over long distances? Indeed, in embryonic stem cells (ESCs), Mediator subunits (MED1 and MED12) co-localize with cohesin at enhancers and promoters, and cohesin is necessary for loop formation between them[25]. Other studies showed that MED1 interacts with GATA1 (ref. 26), a key erythroid transcription factor required for locus control region (LCR) looping to the β-globin gene[27], and the two co-occupy the LCR[28]. Thus, Mediator may coordinate enhancer signalling to the transcription machinery by interacting with enhancer-bound transcription factors and Pol II and serve as a hub for transcriptional regulation by distant enhancers. Another Mediator component, TBP-associated factor 3 (TAF3), interacts directly with CTCF and is recruited to distal sites that are shared by CTCF and cohesin in ESCs[29]. Although it is not clear whether these sites are bona fide enhancers, in at least one example, the distal site loops to a promoter in a TAF3-dependent fashion, and knockdown of TAF3 or CTCF reduces expression of the gene, suggesting that the loop is functional.

Enhancer looping also appears to have a role in Pol II elongation. The LCR and the β-globin looping factor LIM-domain-binding 1 (LDB1) are needed for proper release of Pol II from pausing within the β-globin gene[30,31]. Most recently, the elongation factor ELL3 was found to occupy enhancers in ESCs[32]. The association of ELL3 with the enhancers was required for proper Pol II occupancy at developmentally regulated genes. Both cohesin and Mediator were found to be associated with many ELL3-occupied enhancers and cohesin mediated long-range interactions of an ELL3-occupied enhancer in the homeobox cluster A (*HOXA*) locus. Together, these studies show that enhancers can influence both Pol II initiation and elongation through direct participation of transcription machinery components in looping.

Another means by which enhancers could influence transcription of their target genes is through their own transcription. It has been known for many years that sense and antisense transcripts arise from certain enhancers, although the function of the transcripts was unclear. Is the RNA or the transcription per se important, or is the transcription simply incidental to transcription of a looped gene? Now, genome-wide studies have revealed that enhancers are frequently transcribed into non-coding RNAs of various length, polyadenylation status and strand specificity[33,24,34]. Furthermore, enhancer RNAs (eRNAs) have been used to identify active enhancers, suggesting that enhancer transcription is a part of the enhancer activation process[35]. Transcription of the eRNAs correlated with mRNA synthesis at nearby genes, suggesting an involvement in transcription regulation[33,34]. It seems unlikely that the transcription is a by-product of activation of the target gene, since knockdown of a subset of the eRNAs resulted in decreased gene transcription[34]. This suggests that the RNA itself is required for the enhancer effect, not simply transcription of the non-coding RNA (ncRNA) gene. An intriguing possibility is that eRNAs may have a structural role in establishing or stabilizing enhancer–promoter loops. In fact, new data provide support for this idea[36].

Nevertheless, at this point, the function of enhancer ncRNAs requires considerable further study and validation.

Recent studies document looping interactions between enhancers and promoters on a genome-wide scale. Comprehensive mapping of RNA Pol II-associated long-range interactions in different cell types suggested a structural framework of multi-gene complexes involving close enhancer–promoter interactions to accomplish cell-specific functions[37]. Other studies documented a substantial overlap of CTCF occupancy with the enhancers[38,39], which is consistent with the finding that tissue-specific CTCF sites co-localized significantly with enhancers (50%) in ESCs[12]. These studies of genome-wide enhancer looping have been tied together with eRNAs in a report indicating a significant correlation among gene expression, promoter–enhancer looping and transcription of the enhancers[39]. The picture that emerges is of an ensemble of enhancer–gene interactions that determine a specific cellular transcriptome.

How are these multi-gene complexes organized? The close approximation of active enhancer–gene pairs and similarly regulated genes fits well with the concept of transcription factories that are focal concentrations of RNA Pol II. The coordinately regulated α- and β-globin genes were already known to occupy the same factory much more frequently than they do different ones[40]. This now seems likely to be a generalizable phenomenon. Can a connection be drawn between transcription factory residence and loops between enhancers and promoters? Enhancer loops might serve to deliver the activated gene to a transcription factory. In the absence of the LCR or after reduction of the looping factor LDB1, β-globin loci fail to migrate to factories, implicating enhancer loop formation as a prerequisite[41,31]. However, other scenarios can be envisioned, and this question remains to be rigorously addressed.

Future work may broaden the perspective; however, thus far, mechanistic insights into how enhancers bring about gene expression all invoke looping. In some cases, looping can directly involve components of the transcriptional machinery. Moreover, looping may be influenced by enhancer transcription. Finally, enhancer looping on a genome-wide scale may organize active regions of the genome and may determine the destiny of certain genes for transcription factories. Is enhancer looping sufficient for gene activation? Forcing an enhancer–gene loop in the absence of the normal transcription regulators in the β-globin locus was sufficient to activate transcription at least partially, supporting the idea that enhancer looping causally underlies the transcriptional change[18]. How looping relates to transcription factory residence is an intriguing question. A single-cell technology for determining interaction frequency between sites in chromatin, comparable to the resolution obtained using FISH, would be a substantial advance. Other pressing needs for the future are to determine in an unbiased way the proteins underlying nuclear enhancer–promoter loop organization and to uncover how movement in the nucleus orders the landscape for gene expression.

Q *How do mutations and variants in enhancers influence human disease?*

Marcelo A. Nobrega. About 85% of human DNA under evolutionary constraint corresponds to non-protein-coding sequences[42], a sizeable fraction of which constitutes *cis*-regulatory elements. It is not surprising, thus, that genetic variation within these regulatory sequences has the potential of resulting in phenotypic variation and underlies the aetiology of human diseases. Early examples of altered gene regulation as a mechanism of human diseases emerged over three decades ago, with the demonstration that translocations in the β-globin gene cluster result in thalassaemias. In the absence of mutations in the globin genes, the disease emerged as a consequence of the disruption in the linear relationship between the globin genes and their distant *cis*-regulatory elements[43].

Over the past decade, genomic sequencing efforts confirmed these predictions and afforded a better understanding of the pervasiveness of mutations in distant *cis*-regulatory elements — the vast majority of which are enhancers — underlying human diseases[44]. The picture that has gradually emerged from these studies is that regulatory mutations result in both Mendelian and complex disease traits, that their frequency spectra range from rare to common and that their phenotypic effects range from small to large. However, the functional characterization of putative disease-causing regulatory mutations remains an important challenge, and most mechanistic demonstrations resort to experimental strategies that involve large amounts of labour, cost and time.

Genetic variation in distant enhancers has been linked to several human Mendelian disorders. In an early demonstration of this, mutations in an enhancer controlling the expression of sonic hedgehog (*SHH*) from a megabase away was shown to result in pre-axial polydactyly in families. This phenotype is shared with patients carrying a chromosomal translocation that removes this enhancer from the general vicinity of *SHH*[45]. However, the phenotypic impact of mutations in enhancers may vary substantially from that of protein-coding mutations, even if both are connected to the same gene. Mutations in enhancers are largely limited to *cis* effects on transcription, whereas those in protein-coding sequences may alter broader aspects of gene expression, such as RNA processing and stability, protein folding, and so on[46].

Another central distinction between the impact of coding and non-coding mutations relates to the modularity of distant enhancers: each enhancer of a gene is responsible for a subset of the quantitative, temporal and spatial expression of that corresponding gene. As an example, coding mutations in *TBX5* — a gene involved in heart and forelimb development — results in Holt–Oram syndrome, which is characterized by cardiac and forelimb malformations. Smemo *et al.*[47] showed that the enhancers regulating the cardiac expression of *TBX5* do not regulate limb development and that mutations in these enhancers result in cardiac but not limb malformations, effectively decoupling the heart–limb phenotype usually associated with *TBX5* coding mutations.

While most regulatory mutations leading to disease that have thus far been characterized disrupt pre-existing enhancers, gain-of-function mutations are also likely to participate in disease processes. De Gobbi *et al.*[48] demonstrated how a non-coding variant that segregates in Melanesians in an otherwise non-functional, anonymous stretch of DNA fortuitously creates a functional *cis*-regulatory sequence, resulting in the spurious activation of α-globin

genes and the consequent onset of α-thalassaemia in affected individuals. Thus, the mutational space of non-coding sequences, already estimated to be much larger than that of coding sequences, is likely to be an underestimation of the true figure.

The modularity of enhancers and their functional compartmentalization imply that regulatory mutations will often have a lower burden on fitness than will coding mutations and may reach high frequency in populations. As a prelude to understanding how common variation in distal enhancers might underlie the genetic architecture of several complex traits and diseases in humans, Emison *et al.*[49] demonstrated that common mutations in an intronic enhancer of *RET* increase risk to Hirschsprung's disease, which is a multifactorial disorder. The emergence of genome-wide association studies (GWASs) later confirmed this prediction, and a current estimate is that up to 85% of GWAS loci have non-coding variants as the likely causal association for the trait evaluated[44]. These regulatory variants often reach high frequency in populations and are predicted to affect disease risk through small phenotypic effects, contrasting with the large effect Mendelian variants discussed above.

The precise identification of disease-causing regulatory variants within GWAS loci remains an important challenge, especially in terms of the experimental validation of the putative functional effects of these variants. Nevertheless, a number of regulatory variants in enhancers emerging from GWAS hits have been functionally characterized, and several insights have come out of these studies. First, the same variant may have an impact on the risk for more than one disease[50,51]. Second, new mechanisms of disease have been uncovered or confirmed, such as an altered response to inflammatory signalling underlying the risk of coronary artery disease[52]. Third, uncovering the physiological impact of regulatory variants will often necessitate the use of appropriate cell lines and/or animal models, as illustrated by Musunuru *et al.*[53]. Finally, the detailed characterization of a genetic or signalling pathway associated with a disease may reorient the target of biological exploration. As an example, the association of *TCF7L2* to type 2 diabetes (T2D) has resulted in a large effort to dissect functionally the mechanism for this association in pancreatic β-cells. However, recent data posit that non-pancreatic actions of TCF7L2 may in fact underlie the increased disease risk to T2D[54–56].

Clear challenges remain to be addressed both in the identification of regulatory variants contributing to human diseases and the experimental interrogation of the impact of those variants on biological processes. New technologies that effectively assay functional variants, teasing out their biological impact in high throughput, will be necessary to replace the laborious and low-throughput experimental strategies used thus far. Extrapolating the notion of mutation burden and aggregate analysis used in exome sequencing to regulatory elements will prove a formidable task, and yet it is likely that the genetic architecture of several common diseases will include various regulatory mutations in multiple enhancers within an individual.

Finally, the almost exclusive focus of regulatory mutations on distal enhancers reflects our inability to assay functionally other types of regulatory elements in the genome. Yet other classes of regulatory elements, such as insulators, repressors and matrix attachment regions, are abundant in the human genome and almost certainly have their function modified by rare

and common genetic variants. The development of new experimental assays to interrogate these elements and their putative allelic variants will contribute to uncovering the genetic mechanisms of several human diseases that have as their molecular base variants in distal *cis*-regulatory sequences.

Q *How important are changes in enhancers for evolution?*

Gill Bejerano. Modern technologies driven by next-generation sequencing, such as ChIP–seq, that reveal all genomic DNA in a particular functional state provide breathtaking snapshots of gene regulation in action. We see large amounts of open chromatin, dynamic domains of histone modifications and many thousands of binding sites for virtually any transcription factor and co-factor under any condition[14]. How many of these biochemical events that we observe actually contribute to gene regulation is an open question. How many of these 'matter' (that is, affect fitness) is even harder to answer. With these caveats in mind, it is still safe to assume that at least 5–10 times as much of the human genome codes for gene regulatory functions (10–20%) as is devoted to coding for the transcripts themselves. How this landscape is exactly divided between enhancers and other gene regulatory regions, such as repressors and insulators, and indeed how many of these elements have multiple roles under different cellular conditions is only starting to unfold. The evidence we have suggests that a large fraction of gene regulatory regions can act as enhancers[4,5]. By virtue of occupying so much genome landscape, enhancers provide a large potential target for evolution.

Genome evolution is driven by mutation and selection. An adaptive genomic mutation can take hold by improving fitness in at least one context that the locus is used in while not perturbing function too much in any other context of its use. Most human genes are expressed in multiple cells and tissues at different times. A gene sequence mutation may perturb the organism in all contexts where the transcript is in use, increasing the likelihood of a detrimental effect. The expression domain of a gene, however, is often the sum of inputs from multiple enhancers (and other *cis*-regulatory elements), each active in only a subset of contexts (for example, ref. 47). Thus an enhancer mutation may affect a smaller subset of functional contexts. If it happens to be beneficial in one context, there are fewer other contexts to reconcile with. This modularity makes enhancers even more likely fodder for evolution[57]. Framed by our growing interest in them, there are still a number of fascinating fundamental questions to be addressed at a number of levels about the contribution of enhancers to evolution.

To understand the mechanisms by which enhancers might contribute to evolution, a good starting point is to ask the question 'how are enhancers "born"?' Despite exciting strides being made in determining the biochemical properties of enhancers, we still lack a deep understanding of enhancer logic. For example, we do not know how small or simple enhancers can be at their birth. The handful of enhancers that have been studied in detail are bound by multiple transcription factors over dozens of bases[58]. We also see numerous co-regulated groups of genes in multiple contexts, but our understanding of enhancer logic and gene networks is not deep enough to know how sequence-constrained the enhancers driving these gene groups must be. The more complex and constrained we presume an enhancer

must be to contribute to fitness, the less likely it is that functional enhancers are to arise *de novo* in neutrally evolving DNA. Duplication and divergence of pre-existing enhancers, including in the context of gene duplication, is an appealing model. Surprisingly, however, the majority of human conserved non-exonic genomic loci does not show such homologies[59]. Britten and Davidson[60] proposed an enticing alternative scenario: by scattering thousands and millions of long stretches of almost identical 'repeats' throughout the genome, mobile elements may greatly increase the probability that portions of a fraction of sequences will mutate into enhancers of similar logic and domains of expression next to previously functionally unlinked genes. Multiple groups have made important strides in providing evidence for this hypothesis, yet the prevalence of this mode of gene network evolution remains unknown[61].

Over a million non-coding genomic loci, nearly a fifth of which significantly overlap mobile elements, are evolving under purifying selection in the human genome[62]. Experiments suggest that most are probably enhancers and related gene regulatory components[4,5]. They mutate slowly[62] and are rarely deleted[63] (but see exceptions[64] below). Some human enhancers are staggeringly old[65,66]. This mode of enhancer evolution provides the strongest bulk argument for enhancer contribution to human phenotype evolution.

Multiple researchers, too many to reference individually, have contributed studies of loci where enhancer modifications probably drive the evolution of specific phenotypes. Tellingly, these studies span bilaterians that range from insects to humans, as well as multiple forms of enhancer mutation, including base-pair substitution and deletion[57,67,68]. They include: mutations in insect species that have driven *Drosophila* spp. body and wing pigmentation and larval trichome formation, and butterfly wing patterning; enhancer lesions leading to pelvic fin loss in fish populations; enhancer losses associated with the human-specific loss of vibrissae and penile spines and with brain expansion; and human-specific mutations that change the expression domain of enhancers near important developmental genes. Regulatory mutations also probably underlie many human population-specific adaptations, such as lactase persistence[69].

How widespread is evolution through changes in enhancers likely to be compared with other mechanisms? Ohno[70] speculated that the major driving force in molecular evolution is gene duplication. Differences in gene copy numbers do dominate the literature: for example, the list of studied differences between humans and chimpanzees[71]. But they dominate it by virtue of ascertainment bias. King and Wilson[72] probably grabbed the bull by the horns. The vast majority of nucleotide differences between any two humans, or between humans and related species, are non-coding. Most of these are probably neutral or nearly neutral. But what fraction of mutations leading to a change in phenotype is gene regulatory?

Studies of the genetic variants that underlie common diseases provide insights that help to address this question. Until recently, we expected nearly all genomic mutations leading to human disease to be coding. Now that we can agnostically assay the genome for disease-associated variants, we see that over 80% of GWAS-associated single-nucleotide polymorphisms (SNPs) are non-coding[44] (a similar fraction to our estimate of regulatory sequence versus coding sequence in the genome) as well as a growing list of individual

regulatory loci contributing to human disease[43]. Disease and adaptation are two sides of the same molecular coin. A multiple-population genome-wide study of marine-to-freshwater adaptation in stickleback fish finds a similar split: over 80% of loci carrying the genomic hallmarks of adaptive evolution are probably regulatory[73].

How should we go about increasing our understanding of the contribution of enhancers to phenotype evolution? Because there are so many of them, because several of them may contribute to any single gene expression pattern and mostly because we poorly understand their logic, enhancer functions are harder to pin down[74]. Technology and experimentation will continue to have major roles in our growing understanding of enhancers. But both the space of all states that the genome of an organism regulates and the number of players involved in its regulation are so large that even the biggest consortium-based projects can only chip away at them for the foreseeable future. With so many functional data sets and so many genomes of humans and related species becoming available[75], this is a wonderful opportunity for young researchers to make their mark in deciphering the logic and secrets of enhancer evolution and genotype-to-phenotype relationships. For the data deluge will be most valuable when it is turned into insights to the real deluge, which is life itself.

Finally, it is important to recall that enhancers are only the most widely assayed of the *cis*-regulatory elements. Repression of gene expression and insulator-mediated partitioning of the genome into gene regulatory domains are probably also important contributors to gene regulation and its evolution. This wealth of gene-regulatory functions, the number of elements in the genome that encodes them and the extent to which they contribute to evolution assure that gene regulation will remain an exciting field for years to come.

Acknowledgements

L.A.P. was supported by US National Human Genome Research Institute grants R01HG003988 and U54HG006997. Research was conducted at the E.O. Lawrence Berkeley National Laboratory and carried out under US Department of Energy Contract DE-AC02-05CH11231, University of California. A.D. acknowledges support for research in her laboratory by the Intramural Program of the US National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) at the National Institutes of Health (NIH). M.A.N. is currently supported by the US National Institutes of Health, grants R01DK093972 and R01HL114010. G.B. thanks members of his laboratory, past and present, for their wisdom and company.

References

1. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009; 461:199–205. [PubMed: 19741700]
2. Mohrs M, et al. Deletion of a coordinate regulator of type 2 cytokine expression in mice. Nature Immunol. 2001; 2:842–847. [PubMed: 11526400]
3. Bejerano G, et al. Ultraconserved elements in the human genome. Science. 2004; 304:1321–1325. [PubMed: 15131266]
4. Pennacchio LA, et al. *In vivo* enhancer analysis of human conserved non-coding sequences. Nature. 2006; 444:499–502. [PubMed: 17086198]
5. Visel A, et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nature Genet. 2008; 40:158–160. [PubMed: 18176564]
6. Schmidt D, et al. Five-vertebrate ChIP–seq reveals the evolutionary dynamics of transcription factor binding. Science. 2010; 328:1036–1040. [PubMed: 20378774]
7. Blow MJ, et al. ChIP–seq identification of weakly conserved heart enhancers. Nature Genet. 2010; 42:806–810. [PubMed: 20729851]

8. Visel A, et al. ChIP–seq accurately predicts tissue-specific activity of enhancers. Nature. 2009; 457:854–858. [PubMed: 19212405]
9. Creyghton MP, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl Acad. Sci. USA. 2010; 107:21931–21936. [PubMed: 21106759]
10. Heintzman ND, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genet. 2007; 39:311–318. [PubMed: 17277777]
11. Dorschner MO, et al. High-throughput localization of functional elements by quantitative chromatin profiling. Nature Methods. 2004; 1:219–225. [PubMed: 15782197]
12. Shen Y, et al. A map of the *cis*-regulatory sequences in the mouse genome. Nature. 2012; 488:116–120. [PubMed: 22763441]
13. Zhu J, et al. Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell. 2013; 152:642–654. [PubMed: 23333102]
14. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
15. May D, et al. Large-scale discovery of enhancers from human heart tissue. Nature Genet. 2011; 44:89–93. [PubMed: 22138689]
16. Williamson I, et al. Anterior–posterior differences in HoxD chromatin topology in limb development. Development. 2012; 139:3157–3167. [PubMed: 22872084]
17. Chubb JR, Boyle S, Perry P, Bickmore WA. Chromatin motion is constrained by association with nuclear compartments in human cells. Curr. Biol. 2002; 12:439–445. [PubMed: 11909528]
18. Deng W, et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell. 2012; 149:1233–1244. [PubMed: 22682246]
19. Becker NA, Peters JP, Maher LJ. Mechanism of promoter repression by Lac repressor–DNA loops. Nucleic Acids Res. 2013; 41:156–166. [PubMed: 23143103]
20. Hammar P, et al. The lac repressor displays facilitated diffusion in living cells. Science. 2012; 336:1595–1598. [PubMed: 22723426]
21. Hatzis P, Talianidis I. Dynamics of enhancer–promoter communication during differentiation-induced gene activation. Mol. Cell. 2002; 10:1467–1477. [PubMed: 12504020]
22. Bulger M, Groudine M. Looping versus linking: toward a model for long-distance gene activation. Genes Dev. 1999; 13:2465–2477. [PubMed: 10521391]
23. Krivega I, Dean A. Enhancer and promoter interactions—long distance calls. Curr. Opin. Genet. Dev. 2012; 22:79–85. [PubMed: 22169023]
24. Koch F, et al. Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. Nature Struct. Mol. Biol. 2011; 18:956–963. [PubMed: 21765417]
25. Kagey MH, et al. Mediator and cohesin connect gene expression and chromatin architecture. Nature. 2010; 467:430–435. [PubMed: 20720539]
26. Stumpf M, et al. The mediator complex functions as a coactivator for GATA-1 in erythropoiesis via subunit Med1/TRAP220. Proc. Natl Acad. Sci. USA. 2006; 103:18504–18509. [PubMed: 17132730]
27. Vakoc CR, et al. Proximity among distant regulatory elements at the β-globin locus requires GATA-1 and FOG-1. Mol. Cell. 2005; 17:453–462. [PubMed: 15694345]
28. Kim S-I, Bultman SJ, Kiefer CM, Dean A, Bresnick EH. BRG1 requirement for long-range interaction of a locus control region with a downstream promoter. Proc. Natl Acad. Sci. USA. 2009; 106:2259–2264. [PubMed: 19171905]
29. Liu Z, Scannell DR, Eisen MB, Tjian R. Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. Cell. 2011; 146:720–731. [PubMed: 21884934]
30. Sawado T, Halow J, Bender MA, Groudine M. The β-*globin* locus control region (LCR) functions primarily by enhancing the transition from transcription initiation to elongation. Genes Dev. 2003; 17:1009–1018. [PubMed: 12672691]
31. Song S-H, et al. Multiple functions of Ldb1 required for β-globin activation during erythroid differentiation. Blood. 2010; 116:2356–2364. [PubMed: 20570862]

32. Lin C, Garruss AS, Luo Z, Guo F, Shilatifard A. The RNA Pol II elongation factor Ell3 marks enhancers in ES cells and primes future gene activation. Cell. 2013; 152:144–156. [PubMed: 23273992]
33. Kim TK, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010; 465:182–187. [PubMed: 20393465]
34. Orom UA, et al. Long noncoding RNAs with enhancer-like function in human cells. Cell. 2010; 143:46–58. [PubMed: 20887892]
35. Wang D, et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature. 2011; 474:390–394. [PubMed: 21572438]
36. Lai F, et al. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature. 2013 Feb 17.
37. Li G, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012; 148:84–98. [PubMed: 22265404]
38. Chepelev I, Wei G, Wangsa D, Tang Q, Zhao K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. Cell Res. 2012; 22:490–503. [PubMed: 22270183]
39. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012; 489:109–113. [PubMed: 22955621]
40. Osborne CS, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. Nature Genet. 2004; 36:1065–1071. [PubMed: 15361872]
41. Ragoczy T, Bender MA, Telling A, Byron R, Groudine M. The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation. 2006; 20:1447–1454.
42. Ward LD, Kellis M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science. 2012; 337:1675–1678. [PubMed: 22956687]
43. Kleinjan DA, Lettice LA. Long-range gene control and genetic disease. Adv. Genet. 2008; 61:339–388. [PubMed: 18282513]
44. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl Acad. Sci. USA. 2009; 106:9362–9367. [PubMed: 19474294]
45. Lettice LA, et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. 2003; 12:1725–1735. [PubMed: 12837695]
46. Sakabe NJ, Savic D, Nobrega MA. Transcriptional enhancers in development and disease. Genome Biol. 2012; 13:238. [PubMed: 22269347]
47. Smemo S, et al. Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. Hum. Mol. Genet. 2012; 21:3255–3263. [PubMed: 22543974]
48. De Gobbi M, et al. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science. 2006; 312:1215–1217. [PubMed: 16728641]
49. Emison ES, et al. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. Nature. 2005; 434:857–863. [PubMed: 15829955]
50. Pomerantz MM, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nature Genet. 2009; 41:882–884. [PubMed: 19561607]
51. Wasserman NF, Aneas I, Nobrega MA. An 8q24 gene desert variant associated with prostate cancer risk confers differential *in vivo* activity to a MYC enhancer. Genome Res. 2010; 20:1191–1197. [PubMed: 20627891]
52. Harismendy O, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. Nature. 2011; 470:264–268. [PubMed: 21307941]
53. Musunuru K, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 2010; 466:714–719. [PubMed: 20686566]
54. Boj SF, et al. Diabetes risk gene and Wnt effector Tcf7l2/TCF4 controls hepatic response to perinatal and adult metabolic demand. Cell. 2012; 151:1595–1607. [PubMed: 23260145]

55. Savic D, Park SY, Bailey KA, Bell GI, Nobrega MA. *In vitro* scan for enhancers at the TCF7L2 locus. Diabetologia. 2012; 56:121–125. [PubMed: 23011354]
56. Savic D, et al. Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism. Genome Res. 2011; 21:1417–1425. [PubMed: 21673050]
57. Carroll SB. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell. 2008; 134:25–36. [PubMed: 18614008]
58. Panne D, Maniatis T, Harrison SC. An atomic model of the interferon-β enhanceosome. Cell. 2007; 129:1111–1123. [PubMed: 17574024]
59. Bejerano G, Haussler D, Blanchette M. Into the heart of darkness: large-scale clustering of human non-coding DNA. Bioinformatics. 2004; 20(Suppl. 1):i40–i48. [PubMed: 15262779]
60. Britten RJ, Davidson EH. Repetitive and nonrepetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q. Rev. Biol. 1971; 46:111–138. [PubMed: 5160087]
61. Rebollo R, Romanish MT, Mager DL. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu. Rev. Genet. 2012; 46:21–42. [PubMed: 22905872]
62. Lindblad-Toh K, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011; 478:476–482. [PubMed: 21993624]
63. McLean C, Bejerano G. Dispensability of mammalian DNA. Genome Res. 2008; 18:1743–1751. [PubMed: 18832441]
64. Hiller M, Schaar BT, Bejerano G. Hundreds of conserved non-coding genomic regions are independently lost in mammals. Nucleic Acids Res. 2012; 40:11463–11476. [PubMed: 23042682]
65. Clarke SL, et al. Human developmental enhancers conserved between deuterostomes and protostomes. PLoS Genet. 2012; 8:e1002852. [PubMed: 22876195]
66. Royo JL, et al. Transphyletic conservation of developmental regulatory state in animal evolution. Proc. Natl Acad. Sci. USA. 2011; 108:14186–14191. [PubMed: 21844364]
67. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Rev. Genet. 2012; 13:59–69. [PubMed: 22143240]
68. Levine M. Transcriptional enhancers in animal development and evolution. Curr. Biol. 2010; 20:R754–R763. [PubMed: 20833320]
69. Fang L, Ahn JK, Wodziak D, Sibley E. The human lactase persistence-associated SNP-13910*T enables *in vivo* functional persistence of lactase promoter-reporter transgene expression. Hum. Genet. 2012; 131:1153–1159. [PubMed: 22258180]
70. Ohno, S. Evolution by Gene Duplication. Springer-Verlag; 1975.
71. O'Bleness M, Searles VB, Varki A, Gagneux P, Sikela JM. Evolution of genetic and genomic features unique to the human lineage. Nature Rev. Genet. 2012; 13:853–866. [PubMed: 23154808]
72. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science. 1975; 188:107–116. [PubMed: 1090005]
73. Jones FC, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012; 484:55–61. [PubMed: 22481358]
74. Shim S, Kwan KY, Li M, Lefebvre V, Sestan N. *Cis*-regulatory control of corticospinal system development and evolution. Nature. 2012; 486:74–79. [PubMed: 22678282]
75. Hiller M, et al. A 'forward genomics' approach links genotype to phenotype using independent phenotypic losses among related species. Cell Rep. 2012; 2:817–823. [PubMed: 23022484]

The contributors

Len A. Pennacchio is Senior Staff Scientist in the Genomics Division at Lawrence Berkeley National Laboratory (LBNL), Berkeley, California, USA, and Deputy Director of the US Department of Energy (DOE)'s Joint Genome Institute, Walnut Creek, California, USA. He has an extensive background in mammalian genetics and genomics as well as in DNA-sequencing technologies and their application in addressing outstanding issues in both the medical and energy sectors. He received his Ph.D. in 1998 from the Department of Genetics at Stanford University, California, USA, and carried out his postdoctoral work as an Alexander Hollaender Distinguished Fellow at LBNL. He has authored over 100 peer-reviewed publications, and in 2007 he received the Presidential Early Career Award for Scientists and Engineers (PECASE) from the White House for his contributions to the Human Genome Project and understanding mammalian gene regulation *in vivo*.

Wendy Bickmore is the Head of the Chromosomes and Gene Expression Section at the UK Medical Research Council (MRC) Human Genetics Unit, Institute of Genetics and Molecular Medicine (IGMM), at the University of Edinburgh, UK. Her science is based on the principle that genes do not function in isolation from their chromosomal and nuclear context. She has pioneered research into the spatial and temporal organization of the mammalian genome and the implications of this for gene regulation. Her experimental approaches combine cell biology, especially fluorescence microscopy, with genomics, genetics and biochemistry.

Ann Dean is Chief of the Section on Gene Regulation of the Laboratory of Cellular and Developmental Biology, US National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), at the US National Institutes of Health, Bethesda, Maryland, USA. Her work is directed towards understanding how distal regulatory elements, such as enhancers, insulators and silencers, control gene expression during development and differentiation. In particular, she has focused on how chromatin structure is modified by these elements. Her current work is on deciphering the composition and function of protein complexes that underlie long-range regulatory interactions in the genome in mammals. She serves on several editorial boards and is a reviewer for numerous other journals. Work in her laboratory is supported by the Intramural Program of the NIDDK.

Marcelo A. Nobrega is an Associate Professor of Human Genetics at the University of Chicago, Illinois, USA. His scientific interests include the functional characterization of non-coding sequences in the human genome. As a member of the ENCODE Consortium, his laboratory developed technologies that helped in the identification of tissue-specific, distant enhancers and the demonstration of mechanisms by which disease-associated genetic variants within these enhancers increase risk to complex diseases such as prostate cancer, diabetes and cardiovascular diseases. His current research is geared towards the functional annotation of other categories of non-protein coding DNA — such as insulators, repressors and those encoding long non-coding RNAs — and their impact on phenotypic variation and diseases.

Gill Bejerano, Assistant Professor of developmental biology and computer science at Stanford University, California, USA, discovered ultraconservation in 2003 while clustering human non-coding DNA into families. He also studied mobile element co-option and non-coding genome evolution. His laboratory studies human gene regulation. Their widely used GREAT tool annotates *cis*-regulatory data sets for function, and their recent PRISM resource predicts novel transcription factor functions. They study tissue development and are particularly interested in the placenta and neocortex. They discovered the first enhancers conserved between deuterostomes and protostomes, worked on regulatory elements uniquely missing in humans and are currently studying personal genomes. The laboratory also recently developed a novel ‘forward genomics’ approach to link genome and trait evolution.