From remote enhancers to gene regulation: charting the genome’s regulatory landscapes

Orsolya Symmons and François Spitz

Developmental Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Vertebrate genes are characterized by the presence of cis-regulatory elements located at great distances from the genes they control. Alterations of these elements have been implicated in human diseases and evolution, yet little is known about how these elements interact with their surrounding sequences. A recent survey of the mouse genome with a regulatory sensor showed that the regulatory activities of these elements are not organized in a gene-centric manner, but instead are broadly distributed along chromosomes, forming large regulatory landscapes with distinct tissue-specific activities. A large genome-wide collection of expression data from this regulatory sensor revealed some basic principles of this complex genome regulatory architecture, including a substantial interplay between enhancers and other types of activities to modulate gene expression. We discuss the implications of these findings for our understanding of non-coding transcription, and of the possible consequences of structural genomic variations in disease and evolution.

1. Introduction

An essential part of the control of gene expression is achieved at the transcriptional level. This level of regulation integrates the contribution of multiple types of cis-acting genomic elements. Beyond the promoter region, which is in close proximity to the transcriptional start site, the importance of elements located at much farther distances is increasingly being recognized [1–3]. Growing numbers of human genetic conditions have been found to result from mutations, deletions or other alterations of regulatory elements, mostly enhancers, that can be located more than 1 Mb away from the gene they regulate [4–11]. Genome-wide association studies, which analyse the genetic basis of phenotypic variation, have also frequently identified genomic intervals devoid of protein-coding genes as causal regions, further suggesting that variations or changes in gene regulatory elements can have profound physiological effects [12].

For these reasons, defining the regulatory sequences that control gene activity is becoming a key biomedical issue. Recent technological advances are providing an increasing repertoire of methods to identify regulatory elements, particularly enhancers. As illustrated recently by the ENCODE project, chromatin immunoprecipitation (ChIP) for different transcription factors (TFs), chromatin-associated marks and transcription-associated protein complexes enables the cataloguing of sequences with characteristics of regulatory elements [13]. Strikingly, these efforts revealed that in a given cell, a very large number of elements can regulate a given gene. These findings were reinforced by data from chromatin conformation capture (3C) analyses, which showed that promoters and enhancers are engaged in complex interlaced interactions [14,15], indicating that interactions may extensively reshape the individual activities of gene regulatory elements in a non-additive manner. It is therefore important to complement the approaches that deconstruct the genome into its most basic constitutive elements (promoters, enhancers, silencers, insulators), with more holistic approaches that address how the activities of large arrays of cis-regulatory elements are integrated, to ultimately give rise to gene-specific programmes.
Here, we provide a brief review on the evidence that emphasizes the role of remote \textit{cis}-regulatory elements in gene expression and disease. We describe the different methods that can be used to identify and characterize these elements, particularly enhancers. We present and discuss different findings, which demonstrate that the activity of enhancers is not purely determined by their sequence but can also be highly dependent on the surrounding genomic regions. We summarize our own observations obtained with genome regulatory organization mapping with integrated transposons (GROMIT), a strategy, which we developed to investigate the regulatory activities present within the genome [16]. GROMIT revealed the widespread presence of tissue-specific regulatory activities throughout the genome: these activities are distributed along large intervals, forming broad regulatory landscapes, which extend far away from genes. Comparison of the activities displayed within these regulatory landscapes with those determined for isolated single enhancer elements, demonstrated that a substantial part of these latter, potential activities may be silenced or repressed in one way or another. We discuss what these findings tell us about the nature of the genome’s regulatory architecture, as well as their implications for human disease and for the evolution of gene expression.

2. \textit{Cis}-regulatory elements, disease and phenotypic variation

The importance of regulatory effects in disease and more generally in shaping vertebrate phenotypic variation has almost become a paradigm in modern genomics. Early on, the observation that identical or overlapping clinical symptoms could be caused either by mutations in a gene or by chromosomal rearrangements with breakpoints hundreds to thousands of kilobases away from these genes, strongly suggested the presence of remote influences in controlling gene activities. Initially poorly characterized, these influences were collectively defined under the catch-all notion of ‘position effects’ (reviewed in Kleinjan & van Heyningen [17]). It is now known that several of these conditions result from the disruption of a remote regulatory element [7,8,18,19], sometimes present more than 1 Mb away from the gene it is controlling. New technologies, such as array-comparative genome hybridization and next-generation sequencing, enable mapping of genomic structural variations in patients and in normal populations with greater resolution (reviewed in Alkan et al. [20]). This facilitates the identification and analysis of cases with potential ‘position effects’, including hitherto difficult-to-detect genomic changes. Indeed, thanks to these developments, microduplications that encompass regulatory elements have been revealed as a substantial cause of human developmental malformations [21–25].

Beyond Mendelian diseases, genomic and genetic variations affecting regulatory elements also contribute more broadly to phenotypic diversity: a large fraction of variants associated with disease susceptibility and small-effect quantitative traits resides in intergenic regions [12,26,27]. Observations that \textit{intra-} and \textit{inter-species} phenotypic variation is also commonly due to changes in regulatory regions [28–31] further suggest that modulation of gene function through regulatory innovation or modulation can be a driving force of evolutionary changes.

Because of this important role of regulatory elements in health, disease and evolution, extensive efforts have been applied to identify such elements, and to understand how they contribute to gene expression.

3. Identifying regulatory elements

Broadly, the regulatory elements described to date can be grouped into (proximal) promotors, enhancers, repressors and insulators [32] (figure 1a). Our ability to detect such elements, either computationally or experimentally, has improved markedly over the past decade or so, thanks to the discovery of a number of characteristic features (recently reviewed by Hardison & Taylor [33]). For example, the availability of whole genome sequences has enabled the rapid identification of evolutionarily conserved non-coding sequences, which frequently have regulatory functions [34,35], although some elements, undergoing more rapid turnover, are not amenable to such screens [13,36]. \textit{Cis}-regulatory elements are bound by clustered TFs [37], and TF occupancy is linked to diverse chromatin features that can be ascertained: \textit{cis}-regulatory elements often overlap nucleosome-depleted regions, and are characterized by distinct histone composition, histone modifications and by binding of specific proteins, such as transcription cofactors and chromatin remodelling proteins (figure 1a). Chromatin profiling has therefore become an efficient method to allow genome-wide identification of such sites [33]. Nucleosome-depleted regions can be detected by nucleosome hypersensitivity [38] or FAIRE [39], whereas regions associated with specific histones, histone marks or proteins can be identified by ChIP with appropriate antibodies. Such methods, together with the decreasing cost of sequencing, have recently led to detailed inventories of regulatory elements [13,40–42]. It is noteworthy that the presence of specific marks such as H3K27ac on distant enhancer elements can also help distinguish elements that are active, from elements that may only be in a poised state, thus providing ways to use biochemical profiles to infer biological activities [43,44]. In addition, the development of methods, such as 3C and its derivatives (4C, 5C and HiC), which detect chromosomal regions that physically interact with promoters, has also been useful to identify regulatory elements, particularly distant ones [45,46].

However, it should be emphasized that these approaches exploit indirect properties of enhancers and do not assess them in an operational manner (i.e. whether an enhancer actually contributes to gene expression). Indeed, estimating the proportion of TF binding events or chromatin marks that are truly functional is therefore an important on-going debate in the field, as is the definition of ‘biological function’ [47–49].

\textit{In vivo} reporter assays provide a more functional approach to test individual elements, by assessing their ability to drive gene expression. A frequently used reporter assay consists of cloning a putative enhancer fragment upstream of a reporter gene driven by a promoter. The promoter used is often a small neutral promoter region, with minimal or no activity by itself, but that responds accurately to the input of the adjacent enhancer. Accordingly, the activity of the enhancer is revealed by the expression pattern of the reporter gene (figure 1b). The recent development of massively parallel reporter assays [50,51] offers ways to test the activities of thousands of elements simultaneously, and to dissect individual enhancers by testing the influence of thousands of random mutations on their activity. These high-throughput approaches are opening important avenues, but are currently...
largely restricted to cell lines. They can be applied to in vivo conditions, with some limits: for example, in mice, hydrodynamic tail vein injection of DNA constructs results in episomal uptake of fragments, but primarily in the liver [50]. Thus, despite these new technological developments, getting detailed information about enhancer activity across multiple tissues, developmental stages, in both the proper physiologic and epigenetic contexts, may still be better achieved with enhancer-reporter gene transgenes integrated in one-cell embryos. Importantly, new improved vectors and integration systems (transposons, lentiviruses, integrases [52–55]) may facilitate and improve the efficiency of in vivo integrative transgenesis. Already, systematic in vivo transgenic reporter assays have led to collections of enhancer activities that provide direct and important clues about the nature of tissue-specific regulatory elements [56,57].

These transgenic assays have revealed that a great part of the transcriptional activity of genes can be attributed to the action of autonomous regulatory modules, each in charge of a subset of the overall expression pattern. Their action is often described as additive, and adjacent modules, active in different tissues, do not interfere with each other [58]. However, in essence, these assays are reductionist, and test relatively short pieces of DNA sequences in isolation, which are usually randomly integrated into the genome, and therefore out of their natural genomic context. In several instances, enhancer elements at the endogenous locus do not recapitulate the activity that they showed when tested in enhancer assays [59–61]. The observed differences have been put down to non-additive activity of regulatory elements in their native region, or position effects owing to the genomic context where the element is inserted. It may be worth underlining that such ‘context-dependent’ functions have also been reported by other studies, and probably do not simply reflect technical artefacts of transgenic assays. For example, sequence variation explains only a very small fraction of the differential TF occupancy found in human samples [62], suggesting that distant elements or ‘epigenetic’ factors could contribute to the activity of the same element (see Voss et al. [63]). Detailed studies of some endogenous loci, including their functional dissection through deletions and inversions, highlight the importance of regulatory interactions, either through physical interactions as a chromatin hub or regulatory archipelagos [64,65], or through more complex types of ‘regulatory priming’ of an enhancer by another element [66]. Consequently, it is imperative not
only to catalogue regulatory elements, but also to establish how regulation is achieved mechanistically. Importantly, in most transgenic assays, the genomic element of interest is cloned just next to the reporter promoter, whereas the biological activity of remote enhancers results not only from their recruitment of TFs, but also from their capacity to interact with appropriate target gene(s) in the appropriate tissue or cell type. Some of these aspects of enhancer function can be addressed by testing their activity in the context of yeast or bacterial artificial chromosomes (BACs) [10,67]. Owing to their large size, BACs are generally assumed to represent endogenous regulatory landscapes more accurately, although large, complex landscapes may still not be fully covered within an individual clone.

4. From enhancers to target genes

A vital step to comprehend how gene regulation is achieved mechanistically lies in understanding how the interactions between genes and surrounding regulatory elements are controlled. However, the assignment of target genes to regulatory elements can be ambiguous, because regulatory regions can extend over hundreds of kilobases, and the gene most proximal to an enhancer is not necessarily its target [10,11,68,69]. One approach to predict target genes has been to search for co-occurrence of marks specific for active enhancers and promoters, thereby establishing enhancer–promoter units [42]. Alternatively, current views support the idea that most enhancers are engaged in direct physical interactions with their target gene promoters; these interactions can be detected by 3C or its more high-throughput derivatives (4C, 5C, HiC) [46]. These methods allow the identification of regions physically interacting with promoters [15], or—if combined with ChIP for specific proteins (ChLA–PET)—the detections of interactions between regions bound by those proteins [14,70]. Interestingly though, interactions between distal elements and promoters are not exclusive; instead, promoters as well as enhancers are frequently engaged in multiple interactions [14,15]. Whether these interactions are functionally relevant (e.g. to achieve co-regulation of genes), or simply a consequence of other properties remains unclear. Physical proximity of co-regulated gene loci on different chromosomes has also been observed [71], but these situations most probably represent co-localization to discrete subnuclear domains (transcription factories [72]) optimized for coordinated regulation, rather than trans-regulation through elements on another chromosome. The small number of documented cases where genes are controlled by regulatory elements localized on a different chromosome [73–75] most probably represent exceptional—and sometimes debated—systems. Furthermore, overall interchromosomal interactions have been reported to be rather indiscriminate, with no evidence for their being organized by few, specific regions, and the frequency of interactions correlates strongly with the average distance of a given locus from the edge of a chromosome territory [76,77]. Together, with the paucity of interchromosomal interactions whose frequency reaches the levels of known distant promoter–enhancer pairs [15], this argues that for most genes, the elements that contribute to their expression will be found in cis, yet at distances that could frequently be above hundreds of kilobases. Further evidence for the overwhelming importance of cis-regulation comes from genetic crosses between different mouse strains, which has indicated that potentially more than 90 per cent of gene expression differences can be attributed—either partially or completely— to variants acting in cis [78].

In this context, different systems are known to contribute to the specificity of these often distant enhancer–promoter cis interactions, either through direct tethering in favour of target genes [79], or through insulator sequences that block the ectopic activation of other neighbouring genes [80]. Intriguingly, at the same time, multiple observations suggest that enhancers can act also promiscuously, resulting in activation of neighbouring but biologically irrelevant genes. Such collateral activity has been documented for Lsp, adjacent to the Hoxd cluster [10]. Nme4 associated with the α-globin locus [81] and Igβ in the pituitary [82].

These caveats in our current understanding stem in part from the different approaches used to identify individual elements and to assert their function with frequently indirect readouts. This highlights the need for complementary and alternative approaches that may allow assessing regulatory activity in vivo, and in the endogenous context to add to the accumulated molecular information with more functional insights.

5. Charting regulatory landscapes with GROMIT

To bridge this gap in our understanding, we have recently developed a novel method, GROMIT, that enables us to chart the organization and distribution of regulatory activities along chromosomes, providing an integrated and non-gene-centric view of cis-regulatory activity [16]. The basic principle of GROMIT is to distribute a regulatory sensor throughout the mouse genome, by harnessing the properties of the Sleeping Beauty transposon [83,84]. The regulatory sensor consists of a LacZ reporter gene, driven by a 50 bp long fragment of the human β-globin promoter (figure 1c). This short promoter is essentially neutral: it has no activity by itself, but is very sensitive to endogenous regulatory information, without perturbing endogenous gene expression [16]. Therefore, this system makes it possible to determine the regulatory input acting on a given genomic position, where the transposon in inserted. Because the sensor is incorporated into the genome, it measures the integrated regulatory activity of all elements (activating and repressing) acting on that position. This distinguishes GROMIT from reporter assays, which test individual elements in isolation and at random positions. Importantly, the transposon carrying this sensor gene can be remobilized efficiently in vivo, in a cut-and-paste manner, to generate animals with new insertions. The insertion sites can be precisely mapped, and in contrast to other systems show no integration bias towards particular regions or genomic hallmark regions [16]. This property allows the production of a very large number of insertions, and the study of the expression patterns at those positions (figure 2a). These insertions collectively establish a map of how regulatory influences that control gene expression are distributed.

6. Widespread distribution of tissue-specific inputs

Initial analysis of β-galactosidase stainings of more than 150 insertions, collected at stage E11.5 of embryonic development, revealed that almost 60 per cent of tested locations showed expression [16], regardless of their position relative to genes. An expanded dataset confirmed this initial observation.
The vast majority of insertions showed restricted, tissue-specific expression patterns (figure 2a), with fewer than 5 per cent showing widespread expression. This propensity towards tissue-specific expression from most genomic positions highlighted that regulatory activities are distributed along chromosomes, and not centred towards the vicinity of specific regions such as gene promoters. Importantly, the observed patterns frequently shared striking similarity with the activities of neighbouring enhancers that had been characterized previously or with flanking genes or other insertions. Collectively, these results indicated that GROMIT captures biologically relevant regulatory activities, including those acting far from genes.

First, we took advantage of the VISTA enhancer browser [56] to compare the autonomous activities of individual enhancer elements defined by in vivo enhancer assays with the patterns observed at endogenous genomic positions using GROMIT. On a global level, the comparison showed that the expression patterns captured by our regulatory sensor were overall more complex, with more than 75 per cent of insertions showing activity in more than one tissue, whereas in the reporter assay, more than a half of the enhancers showed reproducible expression in only a single tissue (figure 2b). This broader specificity suggested that—in line with our expectations—the GROMIT sensor generally captured the overlapping activity of more than one enhancer at a given position. These findings were confirmed by case studies of individual loci (two examples shown in figure 3, others in [16,65]).

The relative distribution of tissue-specific activities obtained by GROMIT insertions and VISTA enhancers was quite different (figure 2c). For example, the set of enhancers analysed by VISTA

**Figure 2.** Regulatory activity captured by GROMIT and comparison with enhancer activity. (a) We generated a collection of random, genome-wide insertion sites (left, each red arrow representing an insertion site), which we tested one by one for LacZ expression at E11.5 of embryonic development, revealing pervasive, tissue-specific expression throughout the genome (right, a representative set of obtained expression patterns). A complete list of all insertions and the associated expression patterns is available in the TRACER database (tracerdatabase.embl.de). (b) Comparison of the activity of GROMIT insertions (SB insertions) and the activities documented in the VISTA enhancer browser (VISTA enhancers) [56]. The pie charts show how frequently insertions or enhancers displayed activity in a given number of expression domains (indicated by colours). For the analysis, we removed all insertions from the GROMIT dataset that were clearly part of the same regulatory landscapes (showing similar expression patterns and less than 200 kb apart), and only took into consideration VISTA enhancers from early screens (IDs between 1 and 1290), to avoid bias introduced by tissue-specific p300-bound regions [85]. (c) Distribution of tissue-specific activities between GROMIT insertions (blue bars) and VISTA enhancers (green bars). Tissues are indicated along the x-axis, and the y-axis shows the number of observed insertions/enhancers with activity in a given tissue.
was strongly biased to drive expression in fore-, mid- and hindbrain, whereas only 4 per cent of these enhancers were active in the face. The expression of the regulatory sensor did not show a similar preference for neural tissues, but exposed a frequent regulatory potential for expression in facial tissues (18% of all insertions, counting insertions clustered within 200 kb as one). Possibly, these differences may be purely of technical nature: the VISTA dataset was compiled using an Hsp68 promoter fragment [56], whereas GROMIT uses the $\beta$-globin minimal promoter [16], and although enhancers, by their classical definition, should not show promoter preference, some exceptions have been described [86]. Thus, promoter bias might add to the different distribution of tissue-specific activities. But, these discrepancies could also reflect biological phenomena. For example, the VISTA enhancers used for the analysis were picked from sequences with a high degree of evolutionary conservation. It is known that the degree of sequence conservation can be quite variable for enhancers active in different tissues [36], and therefore the different distributions may mirror the different evolutionary constraints associated with regulatory elements active in different tissues. The observed discrepancies could also arise from tissue-specific differences in the regulatory architecture: if the distribution of particular enhancers differs (e.g. if brain enhancers cluster together), or if their range of action is dissimilar (e.g. brain enhancers have a shorter range of action), then a similar shift would be observed. Further studies will be needed to investigate these questions, but this comparison shows the need of combining diverse approaches to get a more complete picture of gene regulatory mechanisms.

7. From enhancers to regulatory landscapes

Frequently, adjacent insertions (or an insertion-endogenous gene pair) shared extensive similarities in their expression patterns, with several cases where insertions were several hundred kilobases apart (examples in figure 3 and in [16]). The widespread activities and large intervals exposed by these observations are reminiscent of the previously described ‘regulatory landscapes’ (genomic domains where otherwise unrelated genes shared
expression specificities) [10,11], and ‘genomic regulatory blocks’ (regions of conserved synteny between genes and non-coding elements among evolutionary distant relatives) [87]. The data obtained by GROMIT imply that their presence is a pervasive feature, and is not restricted to few loci around key developmental regulatory genes. It is yet unclear, if one or few enhancers with particular long-range properties define large regulatory landscapes, or if the conjugated action of multiple co-interacting regulatory modules dispersed across a large interval determines these co-expression territories, as suggested by recent work on the Lsp–Hoxd interval [65].

Naturally, similar gene expression patterns by themselves cannot be considered proof that a given enhancer is regulating reporter gene expression, or that the same enhancers are regulating the endogenous and the reporter gene. Accordingly, it will be exciting to subject tissues with insertions obtained with GROMIT to other assays, such as chromosome conformation capture, to determine whether the insertion sites physically interact with their putative enhancers, and whether the interaction profile of the endogenous gene and the reporter gene is similar, given their often near-identical expression pattern. Similarly, if the TFs binding a given enhancer are known, ChIP and ChIA–PET will provide interesting avenues to detect whether these proteins mediate direct physical interactions and influence the genomic range and distribution of enhancer activities.

8. From regulatory landscapes to gene expression

As mentioned earlier, the expression patterns displayed by the regulatory sensor often overlap with the expression domains of a neighbouring gene, and in line with this, the expression domains captured by GROMIT were generally a composite of the individual activities of the multiple enhancers that surround it (figure 3).

Yet, at a given genomic position, the reporter gene frequently showed only a subset of the expression domains of the endogenous gene (figure 3), as well as sometimes differences with an immediately adjacent enhancer. As mentioned earlier, these discrepancies can be, in part, technical (promoter bias, sensitivity of in situ probes, etc.). However, adjacent insertions sometimes revealed different subsets of the expression pattern of the flanking endogenous gene, showing that all domains could be captured by our promoter, and the causes for these differences are more intricate [16]. This implies that the regulatory elements that control gene expression may have distinct ranges of action, thereby defining different regulatory landscapes, and ultimately resulting in differential gene expression at different positions within a locus. The observation that GROMIT insertions can report activity of an enhancer across hundreds of kilobases underlines that GROMIT is not ideal for precise identification of enhancer location. However, it enables us to define the range of action of enhancers, the extent of large regulatory landscapes, as well as their boundaries. The distance separating two adjacent insertions showing different expression patterns can sometimes be as short as a few tens of kilobases [16], hinting to the position of possible regulatory insulator elements, which can hardly be identified by direct means.

In addition, detailed side-by-side comparison of reporter and endogenous gene expression, with the enhancer activities obtained by transgenic assays, showed that a substantial subset of autonomous enhancer activities were absent at the intact locus, and accordingly failed to activate the endogenous target gene or a nearby sensor. How this is achieved remains unclear (figure 3c). It is possible that enhancer activity per se is inhibited in certain tissues, for example by direct repressors of enhancer activity, possibly by establishing a repressive chromatin structure, which prevents binding of TFs. Alternatively, the ability of an enhancer to activate a gene or a sensor at a distance may be restricted by the chromosomal conformation of the locus or the interaction profile of that enhancer. Regardless of the mechanistic cause, however, these results imply that long-range regulatory activity is not only prevalent throughout the genome, but that the expression domains determined by enhancers are fine tuned locally by their interplay with other factors involved in gene regulation, resulting in a ‘latent’ potential that can be revealed in another genomic context.

Intriguingly, with few insertions within large gene-deserts, we observed expression of the reporter sensor, which was completely at odds with the activity of flanking endogenous protein-coding genes, defining regulatory landscapes that do not seem to include target genes. It is possible that these regulatory activities control the expression of un-annotated or non-coding genes, such as microRNAs (miRNAs) or long non-coding RNAs (lncRNAs) or even act in trans, towards a different chromosome or at extremely long distance. However, it could also suggest that gene-poor regions may be inhabited by a plethora of elements with tissue-specific regulatory potential, without attributed target genes and only latent biological functions.

9. Outlining the regulatory map of the genome

Taken together, the operational scan of regulatory potential provided by GROMIT, and our prior knowledge of transcriptional regulation revealed some novel principles of the global regulatory architecture of the mouse genome (figure 4a). First, we observed the pervasive presence of extended regulatory landscapes. Within these landscapes, related expression patterns were observed at multiple positions by both our regulatory sensor and the endogenous genes. A comparison with known enhancers revealed that these expression patterns are often the integrated output of multiple regulatory elements. Importantly, an endogenous gene may be associated with overlapping yet distinct landscapes, each one with different tissue specificities and covering different intervals. The subdivision of a single genomic locus into different regulatory landscapes may arise from the relative positions and properties of the regulatory elements that lie within it. In addition to the range of action of individual enhancers, structural constraints and higher-order three-dimensional organization of the genome, including lamin-associated domains [90,91], CTCF-loops [92] or topologically associated domains [93,94] may also influence the formation of these landscapes. Thus, the extended regulatory landscapes observed with GROMIT are probably the functional consequence of chromatin and conformational structures, as well as of the spatial range of enhancers and of their extensive interactions, as described for the Hoxd regulatory archipelago [65].

A future challenge will remain to precisely relate our operational view of regulatory elements and transcriptional output, with the underlying mechanics of gene regulation, and define
the causal and functional relationships between regulatory and structural domains. To obtain such an understanding will require a multi-pronged approach, identifying single elements and their intrinsic activity, generating high-resolution, high-density data from systems such as GROMIT to determine the integrated regulatory output, as well as from biochemical assays to map epigenetic modifications and chromosomal interactions. Such detailed studies will also identify what factors contribute to limit the range of action of enhancers, and what causes the local differences in their activity.

10. Further implications: from regulatory maps to gene expression and phenotype

GROMIT exposed the fact that enhancer activities are not exclusively and selectively targeted to gene promoters, but distributed throughout the genome. This suggests that similar to our regulatory sensor, any type of cryptic promoter may acquire intergenic non-coding transcripts, the existence of which has been widely documented [95]. In the light of our findings, the tissue-specific expression of non-coding transcripts could therefore be a trait that can arise simply from their genomic location, and is not automatically a strong indication of functionality. Consequently, the overlap of expression between lncRNAs and their flanking protein-coding genes may not necessarily indicate functional regulation in cis of the latter by the first [96,97], but simply their common location within the same regulatory landscape (figure 4b). This is not to say that all non-coding RNAs are simply transcriptional noise. But the widespread and promiscuous distribution of tissue-specific regulatory activities revealed by GROMIT indicate that expression specificity may not be sufficient indication of functionality, and that additional experimental evidence is necessary to determine the biological relevance of non-coding transcripts [98].

Figure 4. An integrated view of regulatory landscapes and their implications. (a) The activity of enhancers (ovals) is distributed throughout extended regions (regulatory landscapes, indicated by bars in the same colour as the enhancer), rather than only activating specific promoters. Activities of different enhancers can overlap, giving rise to complex expression patterns. Some enhancer activities are masked within their endogenous environment, giving rise to latent activities (indicated by split blue-white oval). Distinct regulatory landscapes can be separated by insulator-like elements (red double triangle). Because regulatory activities are not targeted towards genes, this results in the widespread presence of tissue-specific expression potential, captured with GROMIT insertions (indicated by schematic of transposon). (b) The pervasive presence of regulatory activities within the regulatory landscapes can contribute to transcription of non-coding genes, such as long non-coding RNAs (lncRNAs, indicated by black arrow), giving rise to their tissue-specific expression patterns (shown by transcript in the same colour as regulatory landscape). Novel genes can easily acquire a specific expression, for example by retrotransposition into an existing regulatory landscape. Thus, a retrogene with no activity by itself (white arrow) will gain expression as part of the regulatory landscape (red—blue arrow). (c) Chromosomal rearrangements can also lead to novel expression patterns. As demonstrated by a schematic deletion of the region overlapping the red triangle, rearrangements can results in alterations of the regulatory landscape, thereby putting genes within the range of action of remote regulatory elements. Such ‘position effects’ can contribute to disease [17,88] and to the evolution of gene expression [89].
On the other hand, the prevalence of tissue-specific regulatory activities within large intergenic domains also constitutes opportunities for evolutionary tinkering: they provide activities that can facilitate the emergence of functional lncRNAs, and explain how retrotransposed genes [99], and evolutionarily young genes such as orphan genes [100] or protogenes [101] can obtain specific expression domains (figure 4b).

Importantly, the local modulation of intrinsic regulatory activities implies that structural changes in the genome, such as deletions, duplications and inversion, could alter this interplay, leading to alterations in the regulatory landscape, through specific gain or loss of expression patterns (i.e. the masking or unmasking of regulatory potential). Whereas the juxtaposition of existing genes in a novel regulatory environment could result in acquisition of new expression patterns (figure 4c). This phenomenon could be the cause of several pathological conditions, where genomic rearrangements ‘move’ genes in new regulatory environments, exposing them to novel regulatory influences, either by ‘adopting’ regulatory influences normally acting on a different gene [10,88,102], or by unmasking ‘latent’ regulatory activities. The hereditary mixed polyposis syndrome caused by a duplication upstream of the GREMLIN gene [103] and pre-axial polydactyly caused by duplication of the limb-enhancer region of SHH [24,25] may illustrate this latter phenomenon well. More generally, such a phenomenon provides a new framework for understanding the phenotypic impact of the widespread structural variation found in the human population [104]. Thus, phenotypes may not only arise directly from the deletion or duplication of regulatory elements, but in part also from shifting the boundaries of existing regulatory landscapes, leading to modulation of gene expression in the vicinity of structural variants, up to few megabases away. Importantly, this is in line with experimental evidence of altered gene expression as a consequence of genomic rearrangements, both in humans and mice.

Thus, ultimately, an integrated map of the regulatory organization of the genome, including the position of enhancers, but also their range of action, their interactions, the location of regulatory boundaries and latent regulatory potentials will be instrumental to translate individual genomic sequence information into phenotypic predictions.

We thank the members of the laboratory for helpful discussions and for contributing to the development of the GROMIT system and data-set, as well as the anonymous reviewers for helpful comments. F.S.’s research is supported by grants from the European Commission-FP7 (Health 223210/CIBISTEM), Human Frontier Science Programme (grant no. RGY0081/2008-C) and the German Research Foundation (DFG-SP1331/3-1). O.S. is supported by a fellowship awarded by the Louis-Jeantet Foundation.

References

36. Blow MJ
37. Kadonaga JT. 2004 Regulation of RNA polymerase II et al
weakly conserved heart enhancers.
Nat. Genet.
transcription by sequence-specific DNA binding elements near key developmental regulators in animals.
Cell
2007 FAIRE (formaldehyde-assisted isolation of
sites in chromatin.
Dev. Biol.
2008 MyoD, a lesson in widespread DNA mechanistic diversity of distal transcription from enhancer binding to developmental control.
Proc. Natl Acad. Sci. USA
2002 Germline transmission and tissue-specific gene dosage on development: overexpression of the PAX6 gene causes severe eye abnormalities.
Proc. Natl Acad. Sci. USA
causes preaxial triphalangeal thumb-polysyndactyly syndrome.
Annu. Rev. Biochem.
Patinozos GP, de Kierm M, de Boer L, Wengel A, Imam AMA, Strouboulis J, de Laat W, Grosveld F. 2004 Multiple interactions between regulatory regions are required to stabilize an active chromatin hub. Genes Dev. 18, 1495 – 1509. (doi:10.1101/gad.289704)