Landscapes and archipelagos: spatial organization of gene regulation in vertebrates

Thomas Montavon[1,2] and Denis Duboule[1,2,3]

[1] National Research Centre *Frontiers in Genetics*, University of Geneva, Geneva, Switzerland
[2] School of Life Sciences, École Polytechnique Fédérale, Lausanne, Switzerland
[3] Department of Genetics and Evolution, University of Geneva, Geneva, Switzerland

Vertebrate genes controlling critical developmental processes are often regulated by complex sets of global enhancer sequences, located at a distance, within neighboring gene deserts. Recent technological advances have made it possible to investigate the spatial organization of these 'regulatory landscapes'. The integration of such datasets with information on chromatin status, transcriptional activity and nuclear localization of these loci, as well as the effects of genetic modifications thereof, may bring a more comprehensive understanding of tissue- and/or stage-specific gene regulation in both normal and pathological contexts. Here, we review the impact of recent technological advances on our understanding of large-scale gene regulation in vertebrates, by focusing on paradigmatic gene loci.

The spatial genome

Control of gene transcription is essential to most cellular functions, particularly in multicellular organisms, in which different cell types must implement specific gene expression programs. This control is achieved largely through the activities of specialized *cis*-acting regulatory sequences, such as enhancers and silencers [1]. Pioneering studies of gene regulation mostly focused on transcription units encoding proteins either ubiquitously expressed ('housekeeping' genes) or restricted to specific differentiated cell types. In both cases, transcriptional activation seems to rely on a limited set of regulatory elements, which were usually found in close proximity of the gene, within a few kilobases (kb) of the start site. As a consequence, classical models often consider regulatory regions as part of a gene, which they see as compact units, independent from one another.

Although this view still stands in many cases, the study of genes encoding developmental regulators has led to a novel paradigm. These genes often display highly pleiotropic functions and complex expression patterns, and hence must integrate various distinct regulatory inputs. Accordingly, such genes are controlled by multiple elements located at great distances from the transcription unit, sometimes within introns of other genes [2–5]. In vertebrates, such long-range regulation can sometimes control the transcription of groups of neighboring genes in a given expression domain over large genomic distances, thus defining a 'regulatory landscape' [6,7].

Interestingly, although complex regulatory modalities have been described in *Drosophila*, involving distant enhancers and/or multiple regulations in *cis* [8], the existence of regulatory landscapes of the order of several hundreds of kilobases has not yet been reported in invertebrate species. This intricate and complex organization of control elements thus seems to be rather specific for vertebrates and may have evolved following the two rounds of genome duplication that accompanied the emergence of this group [9]. This regulatory complexity could also be related to the large fraction of noncoding sequences in the vertebrate genome compared with classical invertebrate models; while mice and humans have an average gene density of approximately one gene every 100 kb [10,11], this density is tenfold higher in *Drosophila* [12]. In *Caenorhabditis elegans*, in which gene expression can be routinely recapitulated using short sequences (approximately a few kb) located upstream of the promoters, gene density is of the order of one gene per 5 kb [13]. In the latter two cases, the evolution of potent enhancer sequences at a distance would likely induce deleterious side effects.

Considerable efforts have been devoted recently to identify regulatory elements via high-throughput methodologies, and it appears that, in the human genome, candidate control sequences largely outnumber genes [14,15]. In parallel, technological developments in the analysis of chromatin organization in the nucleus make it possible to map interactions between genes and regulatory elements [16]. Here, we survey these novel technologies and describe some of their contributions to our understanding of large-scale gene regulation. We illustrate a few conceptual advances using selected genetic loci and discuss the relevance of such regulatory mechanisms to the understanding of both the evolution of the regulatory genome and the cause of some human genetic disorders.

Large-scale approaches

Enhancers represent the largest class of distal regulatory elements reported to date. The term 'enhancer' defines the

Corresponding author: Duboule, D. (Denis.Duboule@unige.ch), (Denis.Duboule@epfl.ch).

Keywords: transcription; genome organization; chromatin loops; enhancers; long-range regulation.

0962-8924/$ – see front matter © 2012 Elsevier Ltd. All rights reserved. http://dx.doi.org/10.1016/j.tcb.2012.04.003

Box 1. Genomic approaches to gene regulation

Chromatin immunoprecipitation (ChIP) maps genomic sequences bound by specific proteins or associated with chromatin modifications, such as histone post-translational modifications. ChIP involves chromatin crosslinking followed by immunoprecipitation of protein–DNA complexes using specific antibodies. DNA is analyzed by quantitative PCR (qPCR) to determine enrichment over candidate sites, by hybridization to microarrays (ChIP–chip) or by deep sequencing (ChIP–Seq) to provide genome-wide location maps [93].

Chromosome conformation capture (3C) techniques measure the frequency with which genomic loci are in close proximity within the nucleus. 3C relies on the fixation of chromatin conformations by crosslinking, followed by digestion with a restriction enzyme and intramolecular re-ligation, resulting in the formation of ligation products between sequences close to one another at the time of fixation. This yields a library representing the sum of DNA–DNA interactions over the cell population used as a starting material [16,45]. In 3C protocols, specific interactions between candidate sequences are analyzed by qPCR.

4C (3C-on-chip or circular 3C) allows genome-wide identification of sequences contacting a locus of interest (or 'viewpoint') after selective amplification of these regions by inverse PCR performed on a 3C library [34]. The interacting sequences are identified by hybridization to microarrays (4C) or by deep sequencing (3C–seq or 4C–seq).

5C (3C carbon copy) interrogates mutual interactions between many loci in parallel. A fraction of the 3C library is amplified after annealing and ligation of a pool of oligonucleotides specific for each of the investigated restriction fragments. The resulting library is typically analyzed by deep sequencing [16].

Hi–C allows determination of the three-dimensional organization of the full genome. It involves DNA shearing followed by enrichment for the ligation junctions, which are labeled with biotin. These junctions are identified by paired-end deep sequencing [32]. The resolution of this method is limited by the many sequence reads required to measure all of the interactions occurring within the nucleus, particularly for vertebrate genomes.

ChIP–loop and **ChIA–PET** combine 3C and ChIP approaches, adding an immunoprecipitation step between chromatin digestion and re-ligation. This allows for the detection of DNA–DNA interactions associated with the binding of a protein of interest. ChIP–loop detects interactions between candidate sequences using qPCR, whereas ChIA–PET involves paired-end deep sequencing and provides a comprehensive description of these contacts [33].

ability to potentiate the efficiency of transcription of an associated gene, irrespective of promoter orientation [1]. Functional tests are relatively straightforward and typically involve synthetic assays in which a candidate sequence, isolated from its endogenous genomic context, is tested for its ability to activate a reporter gene either in cultured cells or as a transgene *in vivo*. Early attempts to identify enhancers genome-wide used DNA sequence comparisons, with the assumption that functionally important regulatory sequences should be conserved during evolution. Although a significant fraction of conserved noncoding elements (CNEs) do indeed display enhancer activity [17,18], qualitative and quantitative aspects of sequence conservation are not by themselves predictive of any specific function. In addition, enhancer elements cannot always be detected using available approaches to assess DNA sequence conservation [14,19,20].

It is well accepted that regulatory elements are recognized by combinations of transcription factors. These factors in turn recruit various cofactors such as histone modifiers or chromatin remodeling complexes, which participate in the transcriptional activation of a target gene [21]. The development of chromatin immunoprecipitation (ChIP) techniques, coupled with either hybridization to oligonucleotide arrays (ChIP–chip) or deep sequencing (ChIP–seq, Box 1), has allowed large-scale mapping of bound DNA sequences. In this way, the histone acetyltransferase (HAT) p300 was found at thousands of regions distant to known promoters, in cell-type specific patterns [15,22]. While p300 binding sites are often predictive of tissue-specific enhancer activity in a transgenic assay, even in the absence of obvious evolutionary conservation [19,23], the presence of other HATs may label distinct subsets of enhancers [24].

Genome-wide mapping of histone modifications has also identified specific 'chromatin signatures' associated with enhancers, such as high levels of monomethylation at lysine 4 of the histone H3 tail and low levels of trimethylation of the same residue. DNA segments displaying such signatures can promote transcriptional activation in cell culture assays [22], whereas acetylation of lysine 27 was recently used to distinguish between 'active' and 'poised' enhancer sequences [25,26]. Also, the recruitment of RNA polymerase II (RNAPII) can be observed at a subset of potential enhancers [27,28]. Together, these studies help define a molecular blueprint for regulatory elements and identified tens of thousands of candidate distal enhancers, most displaying some cell-type specificity. However, the genuine functions of these potential regulators in their endogenous contexts remain to be addressed [29].

Conformation studies

Although these various approaches are instrumental in characterizing the regulatory genome and its various implementations in both a stage- and tissue-specific manner, they contribute little to the formal identification of which target genes interact with defined enhancers, because the former can be located at large distances from their control elements, sometimes intermingled with non-target gene loci [30]. This is taken into account by a prominent model of long-range gene activation, which implies a direct physical association between regulatory elements and target promoters, via the formation of chromatin loops [21]. Such an interaction may trigger enhancer-bound factors to interact directly with target promoters. Models involving chromatin loops have gained considerable support with the development of chromosome conformation capture (3C) technologies and variants thereof [16], which provide an estimate of the frequencies of specific DNA–DNA contacts within the nucleus (Box 1 and Figure 1). Using this approach, enhancer–promoter contacts were observed at several loci, suggesting that chromatin looping is a widespread mechanism of action for distal enhancers [31], although alternative mechanisms have been discussed [21].

In these initial studies, however, only a few candidate interactions could be assessed by 3C and hence *a priori* knowledge of which pairs of sequences were likely to interact was required; for example, previously identified regulatory elements and their target promoters. Modifications of the 3C protocol overcame this limitation and lead

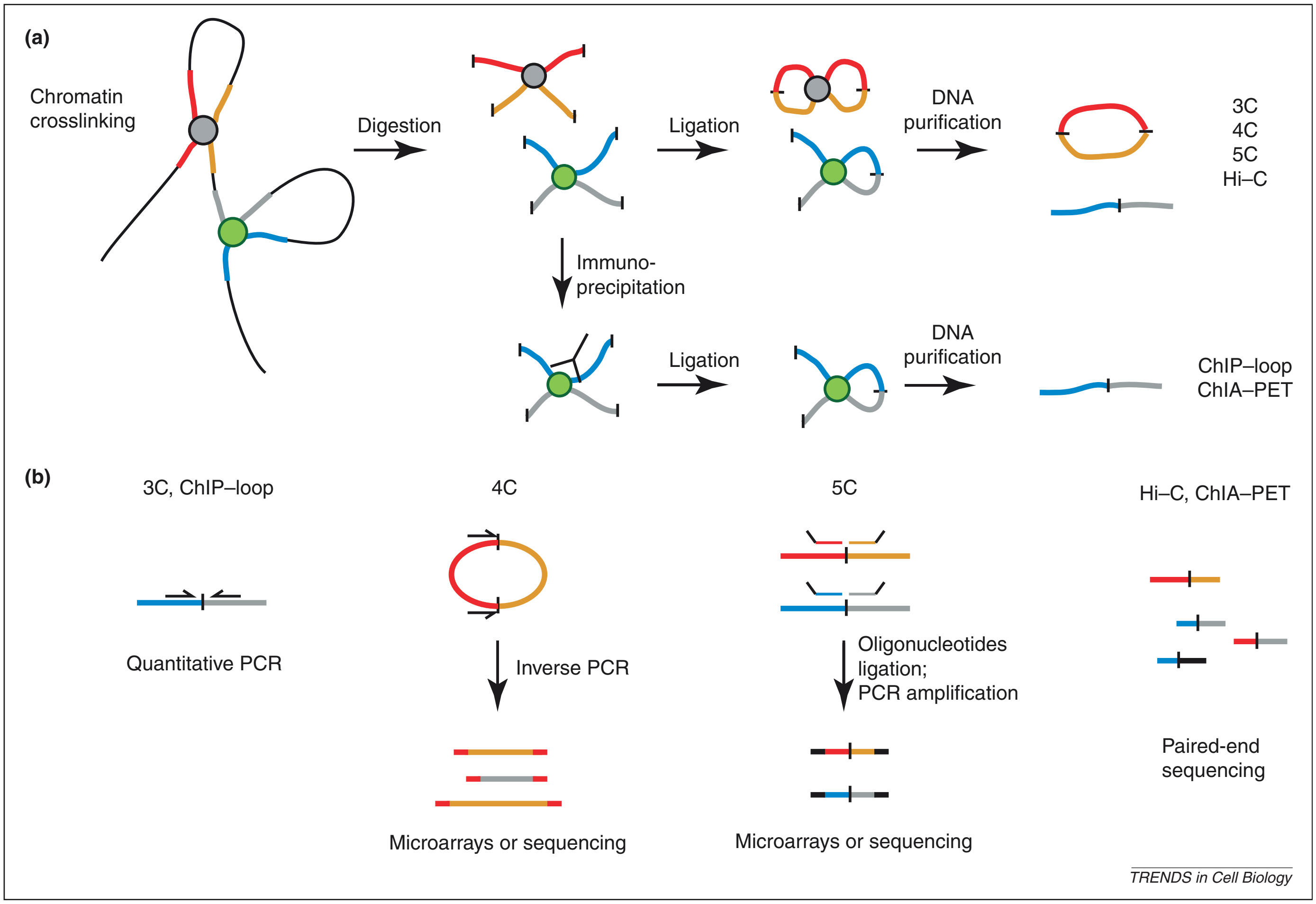

Figure 1. Chromosome conformation capture (3C) approaches. (**a**) General outline of the 3C strategy. Crosslinked chromatin is digested with a restriction enzyme and the restriction fragments are ligated together. The abundance of a given ligation junction in the resulting 3C library is related to the frequency with which the corresponding sequences contact each other within the nucleus. ChIP–loop and ChIA–PET involve an immunoprecipitation step to enrich for sequences bound by a protein of interest. (**b**) Various detection approaches allow visualization of the interactions between candidate sequences (3C), identification of all sequences contacting a locus of interest (4C) or mapping of mutual interactions, either between subsets of the library (5C) or within a complete genome (Hi–C).

either to genome-wide identification of all sequences interacting with a locus of interest (4C) or to the analysis of mutual interactions between many sites in parallel (5C) [16]. More recently, the Hi–C method was developed, which provides a view of chromatin interactions across a complete genome, although at a lower resolution [32]. Finally, approaches like the ChIA–PET integrate ChIP and 3C technologies to identify chromatin interactions associated with specific *trans*-acting factors [33].

The implementation of these methods has revealed that genes often establish complex patterns of contacts, which can involve sequences located several megabases away [34,35]. Chromatin appears to be organized in relatively compact local domains, wherein genes and regulatory regions are spatially clustered [36]. On a broader scale, active and inactive loci segregate into distinct compartments in the nucleus [32,34], which may reflect the recruitment of active genes into transcription factories (i.e. nuclear foci enriched for active RNAPII [37]). The direct visualization of the relative positions of loci via microscopic approaches, such as fluorescent *in situ* hybridization (FISH), complements these biochemical strategies. In the latter case, although the current resolution hardly allows investigation of the details of chromatin conformation within most loci, it can yield insights into the relations between gene expression and localization relative to diverse nuclear landmarks such as chromosome territories (CT) or the nuclear periphery [38].

The integration of conformation studies with a comprehensive identification of regulatory elements will be decisive in the definition of regulatory landscapes and their underlying large-scale mechanisms. Mapping chromatin conformations is indeed insufficient to reveal the role of specific long-range contacts, because similar associations participate in transcriptional repression, and hence repressed loci can also be clustered in the nucleus [39,40]. Enrichment for transcriptionally active chromatin structures, for instance by immunoprecipitation with RNAPII-specific antibodies, can partially overcome this limitation [41,42], yet these approaches still fall short in addressing the functional requirement of the identified partners. In the following sections, we discuss a few selected gene loci where such approaches have been combined with a functional analysis of the regulation at work.

α- and β-Globin loci

Chromatin looping was first documented in the context of the β-globin gene cluster. β-Globin genes are under the control of a major regulatory element, the locus control region (LCR), which is located approximately 50 kb

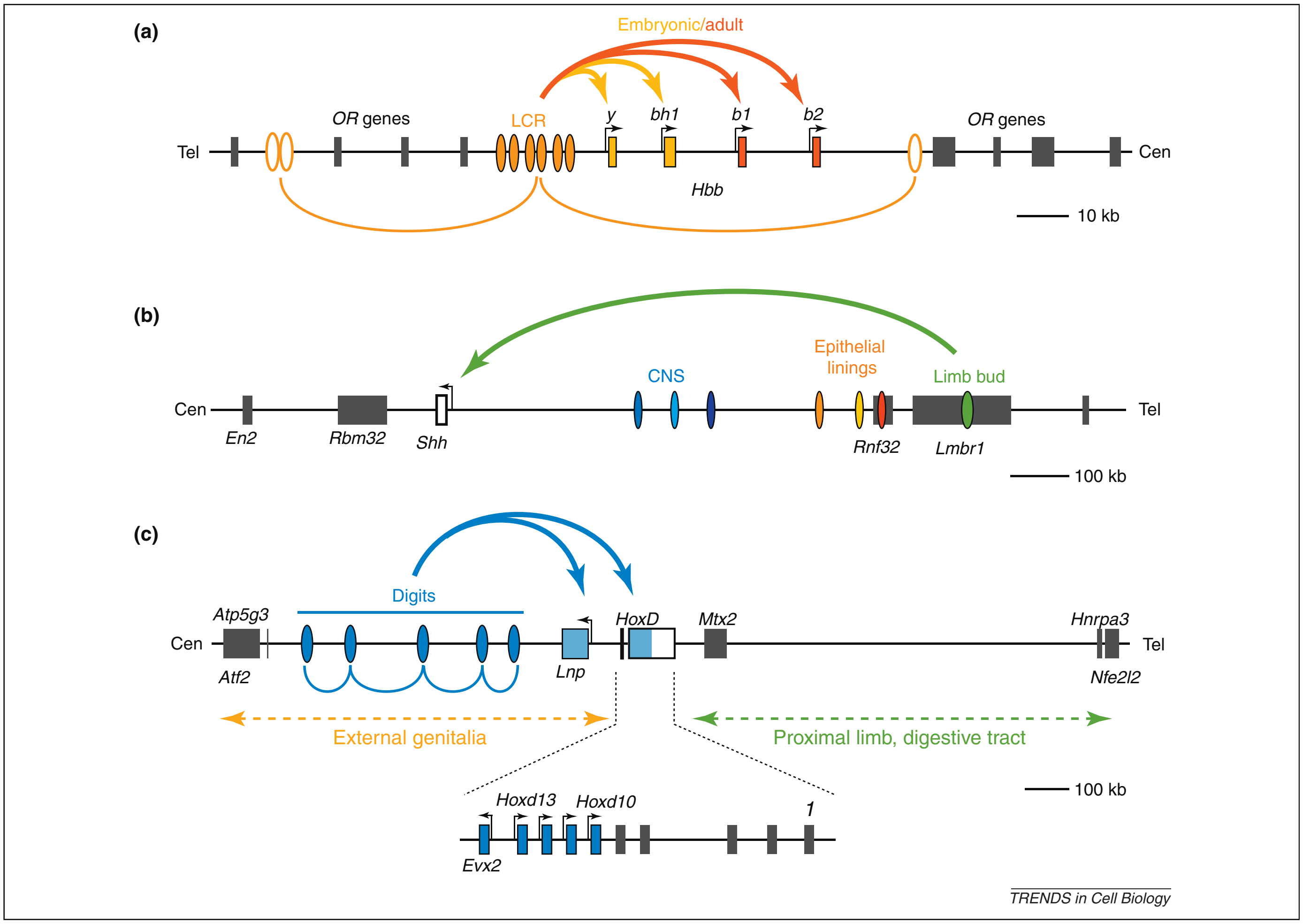

Figure 2. Long-range regulation at selected genetic loci. Genes are represented by rectangles and regulatory elements by ovals. Arrows indicate interactions controlling gene activation; curved lines without arrowheads represent physical associations with unknown functional consequences. Regulations occurring in different tissues or cell types are depicted using different colors. Grey boxes represent genes that are not affected by the described long-range regulations. Note the different scales used for each panel. (**a**) β-Globin (*Hbb*) locus. The locus control region (LCR) contacts and activates either embryonic or adult globin genes at different developmental stages in erythrocytes. Distal sites (open ovals) contact the LCR in both erythroid progenitors and mature erythrocytes (orange lines), yet these sequences are not required for gene activation. (**b**) *Sonic hedgehog* (*Shh*) locus. Candidate enhancers located within the upstream gene desert recapitulate *Shh* expression in specific regions of the central nervous system (CNS) or epithelial linings. An enhancer located within *Lmbr1* contacts *Shh* in the developing limb bud and is required for its expression in this structure. (**c**) The *HoxD* regulatory archipelago. An array of regulatory 'islands' dispersed within the centromeric gene desert coordinately activates *Hoxd13*–*Hoxd10*, as well as *Lnp* and *Evx2* transcription in developing digits. These multiple elements are brought into the vicinity of the *HoxD* cluster in developing digits and each contribute, in a partially redundant manner, to gene activation. Global regulation controlling *Hoxd* genes in different structures relies on control elements located on either side of the gene cluster.

upstream and is necessary for efficient globin transcription [43]. Early 3C studies indicated that the LCR is in close physical proximity to the active globin promoters in erythroid cells, with the intervening DNA looping out [44,45]. Interactions are dynamic, because the LCR selectively contacts the embryonic or adult genes at different developmental stages, and do not occur in cell lineages in which globin genes are inactive (Figure 2a).

Besides its target promoters, the β-globin LCR also contacts distal DNase I hypersensitive (HS) sites and, based on mutual interactions between these sequences, it was proposed that they cluster into an 'active chromatin hub' associated with globin transcription [45]. These distal HS sites are embedded in an array of olfactory receptor genes that are not expressed in the erythroid lineage and do not participate in these interactions. The association between the LCR and upstream sites can already be seen in erythroid progenitors (i.e. cells that do not yet express globin genes), thus forming a poised structure that becomes fully active upon erythroid differentiation. However, some of these contacts may not be critical for transcriptional activation, because deletions of distal sites had no obvious impact on globin expression [43].

Similar chromatin loops have been observed between α-globin genes and their cognate LCR [46,47]. In a recent study, 5C was used to generate a comprehensive interaction map covering a 500 kb region including the α-globin locus. Three-dimensional reconstruction suggested that this domain adopts cell-type specific conformations referred to as 'chromatin globules', where active genes and their regulatory elements cluster towards the core of the structure. In this model, silent chromatin is found at more peripheral locations [36]. Within the nucleus, both α- and β-globin genes preferentially associate with other genes regulated by the same transcription factors, such as *Klf1* [42]. The functional significance of these associations in *trans* remains elusive, but some results suggest they may influence gene expression. A β-globin LCR integrated at an unrelated genomic locus can indeed contact and activate β-globin genes in *trans*, yet in only a small fraction of cells [48].

Sonic hedgehog

The studies mentioned above are concerned with genes whose transcription is required in a single specific cell lineage and hence the long-range mechanisms at work are all involved in this particular task. By contrast, developmental genes with large pleiotropic effects are transcribed in various embryonic structures and at different developmental stages. An example is the *Sonic hedgehog* (*Shh*) gene, which encodes a signaling protein essential for developmental patterning. A transgenic screen identified several long-range enhancers within a large gene desert extending upstream of the *Shh* promoter. When isolated as transgenes, these enhancers could recapitulate various aspects of *Shh* transcription in the central nervous system (CNS) [49]. In the developing limb bud, the expression of *Shh* relies on the activity of another element (ZRS) located almost 1 Mb upstream, beyond the gene desert and within the intron of the *Lmbr1* gene [3]. In addition, three conserved sequences recapitulate *Shh* expression in the epithelial linings of the oral cavity and gastrointestinal tract, with different regional specificities [50]. Short deletions (1 kb) of either the limb or the pharynx elements are sufficient to abolish *Shh* transcription in the corresponding embryonic structures [50,51]. Therefore, *Shh* seems to be controlled by an array of regulatory elements, each dedicated to a specific aspect of its complex expression pattern (Figure 2b).

The spatial conformation of this locus was examined during limb development using both 3C and FISH [30]. Chromatin looping brings the ZRS into the vicinity of *Shh* when limb bud cells are examined, but not into other structures where *Shh* is either silent or expressed under a different control, such as in the CNS. This chromatin loop is observed in a minority of cells both on the posterior part of the limb bud, where *Shh* is active, and in the anterior part, where it is silent. Actively transcribed copies of the gene are found in the vicinity of the enhancer, suggesting that transient associations may trigger transcriptional pulses. Surprisingly, although deletion of the enhancer abrogates *Shh* transcription in budding limbs, it does not affect the conformation of the locus, indicating that looping and transcriptional activation are controlled by different elements. Movement of the *Shh* locus out of its chromosome territory, however, occurs specifically in posterior limb bud cells and requires the presence of the enhancer sequence [30].

Hox gene clusters and regulatory archipelagos

Hox genes encode transcription factors essential for patterning the animal body plan. In mammals, 39 *Hox* genes are grouped into four genomic clusters (*HoxA* to *HoxD*) located on different chromosomes and with similar structural organization. Genes are transcribed sequentially both in time and along the anterior–posterior embryonic axis following their relative position within each cluster, an ancestral phenomenon referred to as colinearity [52]. Because of this additional level of complexity in transcriptional control, long-range regulation at *Hox* gene clusters has been studied in some detail.

Expression of *Hox* genes is tightly linked to their clustered organization, and their transcriptional induction in cultured cells is accompanied by the decondensation of these clusters, as detected by FISH or 3C [53,54]. Recent 4C studies on mouse embryos show that each *Hox* cluster forms a single three-dimensional structure in non-expressing tissues. By contrast, in regions where subsets of *Hox* genes are transcribed, active and inactive genes are separated in distinct spatial domains labeled by different chromatin marks [40,55]. Similar interactions between active *Hoxa* genes were observed in human fibroblasts, yet in this case contacts were not scored between inactive loci [56].

In addition to this collinear regulation that is common to all *Hox* loci, particular vertebrate *Hox* gene clusters have also evolved global expression specificities. For example, distinct groups of neighboring *Hox* genes are coordinately transcribed in the developing limbs, the external genitalia or the digestive tract. Extensive genetic analyses at the *HoxD* locus indicate that these various regulatory landscapes are controlled by long-range elements located within two gene deserts containing many CNEs on either side of the cluster [57] (Figure 2c). For instance, a group of *Hoxd* genes (*Hoxd13* to *Hoxd10*) transcribed in developing digits establishes numerous long-range interactions with sequences dispersed within the centromeric gene desert. These sites of contacts are clustered into 'islands' and are broadly decorated with histone marks associated with enhancer sequences. In addition, they can elicit digit-specific transcriptional activation when isolated in transgenic assays [55,58]. A genetic dissection of this 800 kb DNA interval *in vivo* has revealed that multiple elements contribute, in a partially redundant manner, to the transcriptional activation of *Hoxd* genes in digits. This complex 'regulatory archipelago' could provide both robustness and flexibility to the expression of *Hoxd* genes in digits, a situation somewhat reminiscent of the 'shadow enhancers' described in *Drosophila* [59–62].

Conversely, other *Hoxd* genes, which are not transcribed in digits, preferentially contact sequences located within the telomeric desert on the other side of the gene cluster. Interestingly, some of these long-range interactions are also observed in tissues in which the entire *HoxD* cluster is silent, such as in the developing forebrain, as if the necessary structure controlling gene transcription in digits was already partially preformed [55]. This suggests that the subsequent recruitment of transcription factors may merely trigger the transition from a poised to an active conformation, rather than organizing an entirely new regulatory context.

Gene deserts, regulatory landscapes and genome organization

Both *Shh* and *HoxD* regulatory landscapes are associated with gene deserts, and additional evidence suggests that, rather than being a coincidence, this may reflect a recurrent feature of long-range regulation. Many gene deserts are indeed associated with transcription units of particular importance for the control of embryonic development. These deserts are usually maintained throughout vertebrate evolution and tend to contain a range of conserved DNA sequences potentially required for large-scale regulation [63]; hence, they often overlap with ‘genomic regulatory blocks’ (i.e. regions surrounding developmental genes and defined by both syntenic relations with other species and the presence of arrays of conserved elements [64]). Accordingly, candidate enhancers have been isolated from gene deserts flanking many loci with complex developmental regulation [65–69]. Also, genetic variations associated with human diseases often locate into such deserts (see below). Together, this suggests a function for conserved gene deserts as reservoirs of regulatory information.

Such a concentration of regulatory sequences may in turn keep unrelated transcription units away and thus contribute to the evolutionary stability of gene deserts, given the potential deleterious effects of hosting other transcription units in the vicinity [70]. Although some enhancers do indeed display high specificity for their target promoter and can bypass intervening genes (e.g. *Shh*, [30]), others are more promiscuous and can affect various unrelated genes located nearby [6,71], as illustrated by randomly integrated sensor transgenes, which often adopt expression specificities of nearby genes [72]. Large regulatory landscapes controlled by promiscuous enhancers might be an efficient means ensuring coordinated regulation of functionally related genes in a given domain or at a given time [55,69].

Cis-regulatory mutations in human disease

It was recently estimated that up to 40% of genome-wide association studies point to noncoding DNA intervals as sources of pathology [73]. Structural variations in non-coding portions of the human genome, including point mutations, deletions and duplications, as well as rearrangements separating control elements from their target genes, such as translocations [74], are thus often associated with genetic disorders. Various diseases can also be caused by regulatory mutations affecting the same gene. For example, while mutations in the *SHH* limb enhancer cause hand malformation [3], a point mutation in a candidate CNS enhancer of *SHH* leads to holoprosencephaly [75]. Single nucleotide polymorphisms (SNPs) have been mapped within gene deserts, where they sometimes alter candidate enhancer sequences [65,76,77]. In some cases, 3C was used to link a given SNP to a candidate target gene [76,78–80].

Large chromosome rearrangements and copy number variations (CNVs) can also affect gene regulation in humans [74]. For instance, micro-deletions centromeric of the *HOXD* gene cluster, as well as a balanced translocation with a breakpoint within this gene desert, are associated with malformation of hands and feet similar to those caused by mutations in *HOXD13*, suggesting that they alter the regulatory archipelago necessary for proper gene expression in developing digits [81,82]. Likewise, translocations occurring downstream of *PAX6* in individuals with aniridia (absence of the iris) separate this gene from enhancers active during eye development [74,83,84]. Duplications that include the *SHH* limb enhancer cause three-phalangeal thumb syndrome, probably by changing the dose of this protein during limb development [85]. Finally, deletions, translocations and duplications involving the two gene deserts flanking *SOX9* are associated with various developmental disorders [65,86,87]. In most cases, however, the molecular mechanisms linking noncoding variants to pathological situations remain elusive.

Concluding remarks

Over the past few years, tremendous progress in genomic technologies has been achieved, with far-reaching consequences for our understanding of gene regulatory mechanisms. The emerging picture suggests that vertebrate genes of particular importance for developmental processes are often found surrounded by gene deserts and that these regulatory landscapes can span considerable genomic intervals. Interestingly, although complex gene regulation also exists in classical invertebrate models, it does not seem to involve comparable distances. This difference could be related to gene densities and the need to increase pleiotropic functions in vertebrates. The emergence of such new regulatory modalities may have been triggered by the two full genome duplication events that accompanied the vertebrate radiation. Duplications of gene loci may have indeed allowed complex regulatory rearrangements to occur, without being detrimental to the organism.

To elucidate the intricate interactions that exist between genes and the ever-expanding repertoire of putative control elements will require investment in mapping the three-dimensional organization of the genome [88]. Future directions include the understanding of how specific interactions are formed in the nucleus, and advances in microscopy approaches may allow for a more dynamic visualization of these contacts [89], for instance using live-cell microscopy. Also, identification of the *trans*-acting factors mediating DNA looping is only beginning, and various reports point to an involvement of transcription factors, structural proteins, chromatin modifications or noncoding RNA in this process [90–92]. Yet in many cases, the proposed mechanistic models for long-range activation rely on correlations and indirect evidence. Genetic analyses on living organisms will be critical in deciphering the regulatory logic of such complex loci.

Along with a mechanistic understanding of gene regulation, comparisons between regulatory modalities either in different tissues and cell types or among various animal species will allow us to reconstruct their evolutionary histories. Thus, much in the way that comparing gene sequences has contributed to identifying a link between phylogeny and structural variations, comparative regulomics will allow us to trace the hidden origins of our regulatory circuitries and assess the importance of their modifications in the acquisition and evolution of functions.

Acknowledgments

We thank Michael Levine for discussions. We acknowledge funding from the École Polytechnique Fédérale de Lausanne (EPFL), the University of Geneva, the Swiss National Research Council (SNF), the European Research Council (ERC) and the FP7 EU program IDEAL.

References

1 Maston, G.A. *et al.* (2006) Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* 7, 29–59

2 Zuniga, A. *et al.* (2004) Mouse limb deformity mutations disrupt a global control region within the large regulatory landscape required for Gremlin expression. *Genes Dev.* 18, 1553–1564

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

4 Noonan, J.P. and McCallion, A.S. (2010) Genomics of long-range regulatory elements. *Annu. Rev. Genomics Hum. Genet.* 11, 1–23

5 Hill, R.E. and van Heyningen, V. (2008) Long-range control of gene expression. Preface. *Adv. Genet.* 61, xiii–xv

6 Spitz, F. and Duboule, D. (2008) Global control regions and regulatory landscapes in vertebrate development and evolution. *Adv. Genet.* 61, 175–205

7 Spitz, F. *et al.* (2003) A global control region defines a chromosomal regulatory landscape containing the *HoxD* cluster. *Cell* 113, 405–417

8 Maeda, R.K. and Karch, F. (2011) Gene expression in time and space: additive vs hierarchical organization of *cis*-regulatory regions. *Curr. Opin. Genet. Dev.* 21, 187–193

9 Duboule, D. (2007) The rise and fall of *Hox* gene clusters. *Development* 134, 2549–2560

10 Waterston, R.H. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562

11 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

12 Adams, M.D. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195

13 *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282, 2012–2018

14 ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816

15 Heintzman, N. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459, 108–112

16 van Steensel, B. and Dekker, J. (2010) Genomics tools for unraveling chromosome architecture. *Nat. Biotechnol.* 28, 1089–1095

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

18 Visel, A. *et al.* (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.* 40, 158–160

19 Blow, M.J. *et al.* (2010) ChIP–Seq identification of weakly conserved heart enhancers. *Nat. Genet.* 42, 806–810

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

21 Bulger, M. and Groudine, M. (2011) Functional and mechanistic diversity of distal transcription enhancers. *Cell* 144, 327–339

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

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

24 Krebs, A.R. *et al.* (2011) SAGA and ATAC histone acetyl transferase complexes regulate distinct sets of genes and ATAC defines a class of p300-independent enhancers. *Mol. Cell* 44, 410–423

25 Creyghton, M.P. *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.* 107, 21931–21936

26 Rada-Iglesias, A. *et al.* (2011) A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470, 279–283

27 Kim, T-K. *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465, 182–187

28 De Santa, F. *et al.* (2010) A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol.* 8, e1000384

29 Ahituv, N. *et al.* (2007) Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* 5, e234

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

31 Miele, A. and Dekker, J. (2008) Long-range chromosomal interactions and gene regulation. *Mol. Biosyst.* 4, 1046–1057

32 Lieberman-Aiden, E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293

33 Fullwood, M.J. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 462, 58–64

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

35 Splinter, E. *et al.* (2011) The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev.* 25, 1371–1383

36 Baù, D. *et al.* (2011) The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.* 18, 107–114

37 Sutherland, H. and Bickmore, W.A. (2009) Transcription factories: gene expression in unions? *Nat. Rev. Genet.* 10, 457–466

38 Geyer, P.K. *et al.* (2011) Nuclear organization: taking a position on gene expression. *Curr. Opin. Cell Biol.* 23, 354–359

39 Bantignies, F. *et al.* (2011) Polycomb-dependent regulatory contacts between distant *Hox* loci in *Drosophila*. *Cell* 144, 214–226

40 Noordermeer, D. *et al.* (2011) The dynamic architecture of *Hox* gene clusters. *Science* 334, 222–225

41 Li, G. *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148, 84–98

42 Schoenfelder, S. *et al.* (2010) Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.* 42, 53–61

43 Palstra, R.J. *et al.* (2008) Beta-globin regulation and long-range interactions. *Adv. Genet.* 61, 107–142

44 Palstra, R-J. *et al.* (2003) The beta-globin nuclear compartment in development and erythroid differentiation. *Nat. Genet.* 35, 190–194

45 Tolhuis, B. *et al.* (2002) Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol. Cell* 10, 1453–1465

46 Vernimmen, D. *et al.* (2007) Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *EMBO J.* 26, 2041–2051

47 Vernimmen, D. *et al.* (2009) Chromosome looping at the human alpha-globin locus is mediated via the major upstream regulatory element (HS-40). *Blood* 114, 4253–4260

48 Noordermeer, D. *et al.* (2011) Variegated gene expression caused by cell-specific long-range DNA interactions. *Nat. Cell Biol.* 13, 944–951

49 Jeong, Y. *et al.* (2006) A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. *Development* 133, 761–772

50 Sagai, T. *et al.* (2009) A cluster of three long-range enhancers directs regional *Shh* expression in the epithelial linings. *Development* 136, 1665–1674

51 Sagai, T. *et al.* (2005) Elimination of a long-range *cis*-regulatory module causes complete loss of limb-specific *Shh* expression and truncation of the mouse limb. *Development* 132, 797–803

52 Kmita, M. and Duboule, D. (2003) Organizing axes in time and space; 25 years of colinear tinkering. *Science* 301, 331–333

53 Chambeyron, S. and Bickmore, W.A. (2004) Chromatin decondensation and nuclear reorganization of the *HoxB* locus upon induction of transcription. *Genes Dev.* 18, 1119–1130

54 Ferraiuolo, M.A. *et al.* (2010) The three-dimensional architecture of *Hox* cluster silencing. *Nucleic Acids Res.* 38, 7472–7484

55 Montavon, T. *et al.* (2011) A regulatory archipelago controls *hox* genes transcription in digits. *Cell* 147, 1132–1145

56 Wang, K.C. *et al.* (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 472, 120–124

57 Tschopp, P. and Duboule, D. (2011) A genetic approach to the transcriptional regulation of *Hox* gene clusters. *Annu. Rev. Genet.* 45, 145–166

58 Gonzalez, F. *et al.* (2007) Transgenic analysis of *Hoxd* gene regulation during digit development. *Dev. Biol.* 306, 847–859

59 Frankel, N. *et al.* (2010) Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* 466, 490–493

60 Frankel, N. *et al.* (2011) Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. *Nature* 474, 598–603

61 Hong, J-W. *et al.* (2008) Shadow enhancers as a source of evolutionary novelty. *Science* 321, 1314

62 Perry, M.W. *et al.* (2010) Shadow enhancers foster robustness of *Drosophila* gastrulation. *Curr. Biol.* 20, 1562–1567

63 Ovcharenko, I. *et al.* (2005) Evolution and functional classification of vertebrate gene deserts. *Genome Res.* 15, 137–145

64 Kikuta, H. *et al.* (2007) Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* 17, 545–555

65 Benko, S. *et al.* (2009) Highly conserved non-coding elements on either side of *SOX9* associated with Pierre Robin sequence. *Nat. Genet.* 41, 359–364

66 Kokubu, C. *et al.* (2009) A transposon-based chromosomal engineering method to survey a large *cis*-regulatory landscape in mice. *Nat. Genet.* 41, 946–952

67 Navratilova, P. *et al.* (2009) Systematic human/zebrafish comparative identification of *cis*-regulatory activity around vertebrate developmental transcription factor genes. *Dev. Biol.* 327, 526–540

68 Nobrega, M.A. *et al.* (2003) Scanning human gene deserts for long-range enhancers. *Science* 302, 413

69 Tena, J.J. *et al.* (2011) An evolutionarily conserved three-dimensional structure in the vertebrate *Irx* clusters facilitates enhancer sharing and coregulation. *Nat. Commun.* 2, 310

70 De Gobbi, M. *et al.* (2006) A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* 312, 1215–1217

71 Lower, K.M. *et al.* (2009) Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. *Proc. Natl. Acad. Sci. U.S.A.* 106, 21771–21776

72 Ruf, S. *et al.* (2011) Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nat. Genet.* 43, 379–386

73 Visel, A. *et al.* (2009) Genomic views of distant-acting enhancers. *Nature* 461, 199–205

74 Kleinjan, D-J. and Coutinho, P. (2009) *Cis*-ruption mechanisms: disruption of *cis*-regulatory control as a cause of human genetic disease. *Brief. Funct. Genomics Proteomics* 8, 317–332

75 Jeong, Y. *et al.* (2008) Regulation of a remote *Shh* forebrain enhancer by the Six3 homeoprotein. *Nat. Genet.* 40, 1348–1353

76 Harismendy, O. *et al.* (2011) 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature* 470, 264–268

77 Wasserman, N.F. *et al.* (2010) An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. *Genome Res.* 20, 1191–1197

78 Ahmadiyeh, N. *et al.* (2010) 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc. Natl. Acad. Sci. U.S.A.* 107, 9742–9746

79 Pomerantz, M.M. *et al.* (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.* 41, 882–884

80 Sotelo, J. *et al.* (2010) Long-range enhancers on 8q24 regulate c-Myc. *Proc. Natl. Acad. Sci. U.S.A.* 107, 3001–3005

81 Mitter, D. *et al.* (2010) Genotype–phenotype correlation in eight new patients with a deletion encompassing 2q31.1. *Am. J. Med. Genet.* 152A, 1213–1224

82 Dlugaszewska, B. *et al.* (2006) Breakpoints around the *HOXD* cluster result in various limb malformations. *J. Med. Genet.* 43, 111–118

83 Kleinjan, D.A. *et al.* (2006) Long-range downstream enhancers are essential for Pax6 expression. *Dev. Biol.* 299, 563–581

84 McBride, D.J. *et al.* (2011) DNaseI hypersensitivity and ultraconservation reveal novel, interdependent long-range enhancers at the complex Pax6 *cis*-regulatory region. *PLoS ONE* 6, e28616

85 Klopocki, E. *et al.* (2008) A microduplication of the long range *SHH* limb regulator (ZRS) is associated with triphalangeal thumb-polysyndactyly syndrome. *J. Med. Genet.* 45, 370–375

86 Benko, S. *et al.* (2011) Disruption of a long distance regulatory region upstream of *SOX9* in isolated disorders of sex development. *J. Med. Genet.* 48, 825–830

87 Kurth, I. *et al.* (2009) Duplications of noncoding elements 5′ of *SOX9* are associated with brachydactyly-anonychia. *Nat. Genet.* 41, 862–863

88 Sexton, T. *et al.* (2012) Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* 148, 458–472

89 Boyle, S. *et al.* (2011) Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. *Chromosome Res.* 19, 901–909

90 Kagey, M.H. *et al.* (2010) Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 467, 430–435

91 Mishiro, T. *et al.* (2009) Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. *EMBO J.* 28, 1234–1245

92 Orom, U.A. and Shiekhattar, R. (2011) Noncoding RNAs and enhancers: complications of a long-distance relationship. *Trends Genet.* 27, 433–439

93 Park, P.J. (2009) ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10, 669–680