Web-based tools and approaches to study long-range gene regulation in Metazoa

David Fredman, Pär G. Engström and Boris Lenhard

Advance Access publication date 15 July 2009

Abstract

Regulatory elements can affect specific genes from megabase distances, often from within or beyond unrelated neighbouring genes. The task of computational charting of regulatory inputs in the genome can be approached from several directions. Typically, computational identification of putative regulatory elements for a gene of interest requires tools that will aid in estimating the extent of the (potentially vast) genomic region around the gene that is likely to contain regulatory elements, as well as tools for the identification and characterization of individual elements. Conversely, starting from a putative regulatory element or a regulatory variation in a non-coding region, one often wants to associate the regulatory element with the correct target gene(s). The design of tools for these purposes relies on the remarkably high level of sequence conservation of thousands of regulatory enhancers, their strong tendency to cluster around their target genes, as well as a constrained range of functional categories of the corresponding target genes, many of which are developmental regulators. Additional evolutionary information, such as conservation of synteny, and a growing body of functional genomic and epigenomic data are being rapidly added to established and emerging tools for studying developmental regulation and cross-species conservation to provide new functional insights into the roles of these regions. In this article, we give an overview of the functionality available in general purpose and new/specialized web tools for the above tasks, and discuss current and future developments in the field.

Keywords: bioinformatics; comparative genomics; gene regulation; enhancers; conserved non-coding elements; genomic regulatory blocks

INTRODUCTION

A series of studies in recent years has established that many developmental regulatory genes in vertebrate genomes possess a peculiar feature: they are spanned by arrays/clusters of highly conserved non-coding elements (HCNEs) that inhabit their introns, reach deeply into intergenic sequence flanking those genes and even spill into introns of neighbouring, unrelated genes or beyond [1]. A subset of these conserved elements possess enhancer activity in mouse [2, 3], Xenopus [4] and zebrafish [5, 6], and the entire region spanned by them seems to be receptive to their regulatory inputs based on the integration of enhancer trap constructs [7].

An explanatory framework for these loci has been put forward in the form of genomic regulatory blocks (GRBs, reviewed in (Navratilova and Becker) in this issue; see also [1, 8]). Briefly, arrays of HCNEs around target genes define the regions in which long-range regulatory inputs for the target genes reside. Such genomic arrangements were shown to exist in vertebrates [1, 9], nematodes [10]...
and insects [11]. They were shown to underlie the most ancient blocks of conserved synteny in genomes, and genes other than the target gene in those blocks of synteny were termed ‘bystander genes’ for their apparent unresponsiveness to long-range regulation in the region and their ability to relocate outside of the GRB following whole-genome duplication (WGD) [1].

We believe that the GRB concept is central to efficient finding of targets of long-range regulation, for defining the extent of regions containing long-range regulatory elements, and for the reconstruction of evolutionary events at these loci. For this reason, this review will focus on GRB-centric localization and characterization of long-range regulatory inputs using the available web-based resources. It is not meant to exhaustively cover all existing resources, but to focus on the principle and on how to get most useful information out of it. We conclude by pointing the reader to the more powerful tools and skills that will be of essence for future researchers in functional genomics.

BASIC DIVISION OF TASKS
Since GRBs are defined by a high density of HCNEs (please see ref. [12] for synonymous terms and their disambiguation), their characterization involves several distinct tasks:

Finding HCNEs
This task requires that pairwise or multiple genomic sequence alignments of species at suitable evolutionary distance are available, and that at least one of the genomes is sufficiently well annotated for gene content such that the annotation can be used to filter out the coding or all transcribed regions.

Finding GRBs
In the first approximation, GRBs are characterized by clusters of HCNEs, so this task requires a tool that reveals the loci with high densities of HCNEs. Additional information that might be used to estimate the extent of GRBs is the conservation of synteny across large distances.

Finding the target genes of GRBs
The GRB model requires one to distinguish between the genes that respond to long-range regulation (target genes) and those that do not. While there is still no foolproof set of criteria to distinguish the two with absolute certainty, current knowledge allows for fairly reliable educated guesses (see below).

Finding regulatory elements
Ultimately, we want to come down to characterizing the actual regulatory elements and the transcription factors or other trans-acting regulators that act on them.

FINDING HCNEs
All major genome browsers contain tracks that provide information about cross-species conservation. Using the UCSC Genome Browser [13] as an example, the informative tracks are:

- Chained alignments are sets of pairwise alignments between genomic sequences produced by blastz [14] and ‘chained’ together by conserved order (synteny) between the two genomic sequences. They are symmetric between the two aligned genomes and as such useful for resolving the cases where there is no 1:1 correspondence between the two genomic sequences.
- Net alignments are produced by post-processing chained alignments and show the best chain between two genomes for a particular span of sequence, with respect to one of the genomes (the reference genome). They are useful for determining orthologous regions and often clearly reveal genomic rearrangements by showing breaks of synteny as regions where the top (orthologous) alignment changes chromosome (indicated by a colour change). Care is advised, however, when using them within pseudogenes, tandem repeats, lineage-specific gene family clusters and other repeated regions, since orthologous alignment is uncertain in those cases.
- Conservation tracks based on multispecies alignments—the UCSC Genome browser offers multiple alignments produced using multiz [15], with per-base conservation scores calculated with phastCons [16]. These tracks provide a quantitative display of the conservation score, as well as a schematic representation of the multiple alignments. At high magnification, it will show the underlying sequences in the browser itself. Otherwise, clicking on the schematic alignment will retrieve multiple alignment blocks with links to the corresponding locations in genome browsers of other species. These tracks are useful for
inspection of genomic regions around target genes for individual conserved elements. Also, there are pre-computed ‘Most conserved’ sequence features derived from the regions with highest phastCons scores. These regions are annotation-agnostic, so they need to be filtered if e.g. only non-coding conserved sequences are desired. This can be done using the UCSC Table Browser, with further data processing possible in Galaxy (see ‘Beyond web browsers and databases’ below).

The amount of available conservation information (including the number of aligned species), unfortunately, varies greatly for different genomes. The information available for human, mouse and *Drosophila melanogaster* is typically abundant and up to date. For other model organisms, such as zebrafish, multiple alignments and abundance of other comparative genomic features are vastly inferior to those of top-priority species. Yet, for other model organisms such as the sea urchin *Stronglyocentrotus purpuratus* there are still no suitably related species for the comparative genomics of regulatory elements, but low-coverage sequencing of two other sea urchin species is under way (Sea Urchin Genome Project, Human Genome Sequencing Center at Baylor College of Medicine).

Ensembl, the other major genome browser (www.ensembl.org, [17]), offers similar features. The VISTA browser (http://pipeline.lbl.gov, [18]) with its highly responsive interface is especially suitable for quick zooming in on the conserved elements.

On the other hand, purpose-built browsers and data collections provide convenient routes towards identifying conserved sequence, often with added functionality to assess their potential regulatory role and regulatory target genes.

**PRE-EXTRACTED SETS OF HCNES AND THE ASSOCIATED DATABASES**

A number of web-based resources provide HCN sets pre-computed and extracted from different sequence alignments at different conservation criteria and filtered against known coding sequences (Table 1).

VISTA Enhancer Browser [2] is both a repository of data on experimental and bioinformatic characterization of human HCNs tested for enhancer activity in mouse embryos, and a specialized browser based on the VISTA infrastructure that shows the tested element in the genomic context. At the time of this writing, it contained more than a thousand tested elements, out of which about half showed enhancer activity.

CONDOR (the database of COnserved Non-coding Orthologous Regions [19]) contains more than 7000 HCNs conserved between orthologous regions in fugu and tetrapods extracted from multiple alignments. It emphasizes HCNs retained in duplicate in Fugu after the teleost-specific WGD [20], and those retained since the two rounds of WGD at the origin of vertebrates [21]. The HCNs are viewed and searchable by one or more of the developmental target genes in their vicinity.

TFCONES (the database of Transcription Factor Genes & Associated COnserved Noncoding Elements [22]) contains all annotated human, mouse and fugu genes encoding transcription factors and the HCNs located in the regions around them up to the next closest gene in each direction. As many GRBs extend to the introns of neighboring genes and beyond [1, 8, 12], this will miss many of the relevant HCNs. TFCONES can be searched on genes, genomic coordinates or query sequence and provides HCOME information in tabular format and in a genome browser based on GBrowse [23].

CNEviewer [24] is a web resource for retrieving zebrafish HCNs by syntenic proximity to genes, choosing genes by name or by their anatomical expression. Its gene- and anatomy-based interface complements nicely other resources such as Ancora and the ECR browser that are genome-based. It should be noted here that the genes reported as closest to (H)CNs will often be bystander genes, and further examination of the locus using the concept of genomic regulatory blocks is recommended to assign the plausible target gene (see ‘Finding the target gene of long-range regulation’ below).

Ancora (Atlas of Noncoding Conserved Regions in Animals, [25]) provides a wide range of HCNs based on filtered pairwise alignments and HCOME density tracks indicating the amount of non-coding conservation in larger windows. The annotation is available through the Ancora genome browser, which also provides integration with the UCSC, Ensembl and VISTA browsers, and programmatic access.

ECR Browser [26] displays sequence conservation in a custom genome browser that is well suited for exploring conservation in gene-sized regions.
<table>
<thead>
<tr>
<th>Name</th>
<th>Ensembl</th>
<th>UCSC Genome Browser</th>
<th>VISTA</th>
<th>Ancora</th>
<th>CNEviewer</th>
<th>CONDOR</th>
<th>ECR Browser</th>
<th>TFCONES</th>
<th>VISTA Enhancer Browser</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL (http://)</td>
<td><a href="http://www.ensembl.org">www.ensembl.org</a></td>
<td>genome.ensembl.org</td>
<td>genome.igv</td>
<td>ancore.igv</td>
<td>biowinsic.igv</td>
<td>condor.fugu</td>
<td>ecg.browser</td>
<td>tfcones</td>
<td>enhancer.fugu</td>
</tr>
<tr>
<td>Enhancer assay</td>
<td>Mouse (VISTA Enhancer)</td>
<td>Mouse (VISTA Enhancer)</td>
<td>Mouse (VISTA Enhancer)</td>
<td>Mouse (VISTA Enhancer)</td>
<td>Mouse (VISTA Enhancer)</td>
<td>Mouse (VISTA Enhancer)</td>
<td>Mouse (VISTA Enhancer)</td>
<td>Mouse (VISTA Enhancer)</td>
<td>Mouse (VISTA Enhancer)</td>
</tr>
<tr>
<td>Alignment type</td>
<td>Pairwise, Multiple</td>
<td>A majority of assembled metazoan genomes</td>
<td>Multiple</td>
<td>Multiple</td>
<td>Multiple</td>
<td>Multiple</td>
<td>Multiple</td>
<td>Multiple</td>
<td>Mouse (VISTA Enhancer)</td>
</tr>
<tr>
<td>Loci examined</td>
<td>Genome-wide</td>
<td>Genome-wide</td>
<td>Genome-wide</td>
<td>Genome-wide</td>
<td>Genome-wide</td>
<td>Genome-wide</td>
<td>Genome-wide</td>
<td>Genome-wide</td>
<td>Selected HCNEs</td>
</tr>
<tr>
<td>Genome browser</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Download (bulk)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>DAS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Search by gene</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Search by expression</td>
<td>Yes (enhancer assay)</td>
<td>Yes (enhancer assay)</td>
<td>Yes</td>
<td>Yes (gene expression)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table I: Overview of the feature set of the online browsers and databases for identification of conserved sequence covered in this review.
HCNEs can be extracted at arbitrary conservation thresholds and sizes, and their transcription factor binding site (TFBS) content explored via links to rVISTA [27]. Syntenic HCNEs provide anchors for exploring orthologous loci across the different reference genomes available. Pre-calculated sets of conserved elements, classified by their overlap with other annotations, are available for download.

Each of the above resources has a different focus and provides different functionality and different sets of HCNEs, for various sets of species. VISTA Enhancer Browser and CONDOR provide original data on enhancer activity of elements tested in the laboratories that produced the resource. CNEviewer has the genes neighbouring HCNEs annotated using the anatomical ontology from ZFIN [28]. Ancora is unique in that it is the only tool that provides views of conservation on a larger scale. With some effort, all of the general and specialized resources listed above, currently with the exception of CNEviewer, can be integrated to some extent—either by adding custom tracks from other resources to their built in genome browsers, or by viewing their data in other browsers. For example, VISTA elements are directly available in the UCSC and Ensembl browsers for the human genome, Ancora and CONDOR elements can be viewed in Ensembl through the Distributed Annotation System (DAS, [29]), and several resources provide downloadable files that can be directly used as or turned into custom track files.

FINDING GENOMIC REGULATORY BLOCKS

Most GRBs can be located by identifying genomic regions with high HCNE density. While general genome browsers (UCSC, Ensembl) and specialized comparative browsers (VISTA, ECR Browser) are suitable for visualization and extraction of individual conserved elements at high genome magnification, the regions where such elements are most abundant in the genome are not obvious without previous knowledge.

The aforementioned Ancora browser was developed for the purpose of identifying and visualizing GRBs. Ancora’s most prominent feature is a track that shows the density of HCNEs in a large (minimum 10 kb, typically 40–300 kb) sliding window (Figure 1). The HCNE density track is unique to Ancora and particularly informative for viewing large regions, up to entire chromosomes. Ancora currently contains human, mouse, zebrafish (Zv7), D. melanogaster and Caenorhabditis elegans as reference genomes, and several others are being tested at the time of this writing.

There are a number of approaches for estimating the extent of regulatory elements around a target gene. While duplicated elements left over from a whole-genome duplication [30, 31] and synteny conservation between tetrapods and fish [1] are useful estimates, they may be too conservative for some purposes, for example when looking for mammal-specific elements which might have appeared after the root of the tetrapod lineage [32].

The simplest current estimate of the genomic range in which regulatory elements are found around a given gene is based on the identification of (syntenic) arrays of conserved elements in the genome.

FINDING THE TARGET GENE OF LONG-RANGE REGULATION

Paradoxically, given their characteristic pattern of HCNE density, the locations of GRBs are generally easier to determine than are the target genes in those regions. However, several recent observations and advances have revealed what we can safely consider features of GRB target genes:

‘Trans-dev’ function

Most target genes play a role in regulating organism development [1, 6, 9, 33]. Their gene products are not only transcription factors, but can also be signalling molecules (e.g. decapentaplegic, [11]) or cell adhesion molecules involved in CNS development. For genes that have not been well studied, it may be informative to look at their homologs, associated Gene Ontology terms, and other details listed on the gene pages in Ensembl (for Ensembl Genes) and the UCSC browser (for UCSC Known Genes). In particular, transcription factor genes can often be recognized by the presence of predicted DNA binding domains.

Complex expression pattern

Genes with simple expression patterns are less likely to be subject to long-range regulation [1, 21]. Expression patterns can be explored in databases of in situ hybridization images (see below). Where such data is not available, microarray data collections, such
as the ArrayExpress Atlas (http://www.ebi.ac.uk/arrayexpress; [34]) and BioGPS (http://biogps.gnf.org/), may give some guidance.

**Long and/or multiple CpG islands**

Most mammalian promoters coincide with a CpG island, i.e. a region enriched for CpG dinucleotides [35]. CpG dinucleotides are otherwise underrepresented in vertebrate genomes due to mutational processes [36]. Broadly expressed genes and some developmental regulatory genes tend to have CpG-island promoters, whereas tissue-specific genes do not [35]. In addition, target genes are unusual in that they often overlap with CpG islands at their introns and internal exons. Along their length, they tend to overlap more and larger CpG islands than do bystander genes [37]. Interestingly, primate genome comparisons have shown that the CpG dinucleotides in target gene CpG islands tend to be conserved in evolution [38] and appear to play a role in conferring epigenetic state (see ‘Finding elements by experimental binding data’ below). Locations of CpG islands can be viewed alongside genes in the Ensembl, UCSC and Ancora genome browsers.

**Figure 1:**

(A) HCNE density on human chromosome 17 calculated in 300 kb sliding windows over HCNEs detected by comparison to mouse, chicken and zebrafish genomes at similarity thresholds 96% (yellow/light grey), 98% (orange/medium grey) and 100% (red/black) for mouse, 80/90/98% for chicken, and 70/80/90% identity in the zebrafish comparison, over a 50 column sliding window. (B) Detailed view showing genes, CpG islands, human:zebrafish synteny, HCNE density and HCNEs under one of the HCNE density peaks. It is a 790 kb region around the human LHX1 gene defined by a cluster of HCNEs syntenic in all compared species. In the synteny track, boxes indicate aligned segments, lines indicate gaps and labels indicate orientation and position in the zebrafish genome. Of the multiple genes present in the human locus, LHX1 is the gene likely to respond to long-range regulatory elements in the region on the basis of its function (a homeobox gene involved in differentiation and development), its promoter overlapping multiple CpG islands and that on inspection of synteny of HCNEs, and the zebrafish orthologs for the genes in the region, one of the two zebrafish orthologs, lhx1a, is found as in isolation on zebrafish chr15 together with a full complement of the HCNE array.
Distinct chromatin marks
In mammalian cells, target genes tend to be enriched for certain histone modifications and for binding of the repressive Polycomb protein complexes (see below).

Synteny conservation
If the function of a HCNE is to control the expression of a particular gene in cis, the HCNE should be in conserved synteny with that gene in all genomes where the HCNE is conserved. Synteny conservation can therefore be a powerful indicator for determining target genes. Still, interpreting synteny conservation can be far from straightforward, because GRBs often contain additional (bystander) genes that have been retained within the GRB due to selection against rearrangements that break up a HCNE array [1, 11]. Importantly, whole-genome duplication can relieve this selective pressure. The teleost fish underwent a whole-genome duplication following their divergence from the tetrapods, and their genome sequences therefore provide an excellent source of information for distinguishing target and bystander genes [1]. The rearrangements that have occurred in paralogous teleost GRBs following WGD can be retraced by comparing them to their orthologous GRB in the human genome (or another well-assembled tetrapod genome). At present, for analysis across all teleosts this is most easily done using the chained alignment track in the UCSC browser. See also Figure 1 and ref. [25] for examples of how to interpret duplicated GRBs in Ancora. For mammal: fugu comparisons, the CONDOR database [19] provides detailed information on HCNEs present in duplicated trans-dev regions.

Additional criteria for identifying target genes are emerging. We have recently established that target genes tend to have more alternative transcription start sites than bystander genes (possibly extending the number of combinations of enhancer–promoter interactions), and that their alternative first exons tend to be more widely spaced [37]. Based on this growing list of features, it should soon be feasible to develop an automated procedure that, in most of the cases, efficiently identifies target genes among multiple possible genes in a GRB.

FINDING ELEMENTS USING EXPERIMENTAL BINDING DATA
From the first genome-wide studies using chromatin immunoprecipitation (ChIP)-chip and ChIP-seq technologies to study histone methylation, acetylation and binding of transcription factors and co-factors, several observations have been striking. In mouse embryonic stem cells, ‘bivalent’ chromatin domains with overlapping repressive (H3K27me3) and activating (H3K4me3) histone modifications mark the promoters of over 2000 genes [39]. The subset of those domains that are bound by Polycomb repressive complex 1 and 2 associate with developmental regulatory genes and can be predicted by clusters of large CpG islands depleted of activating transcription motifs [40]. Among the genes associated with such epigenetic marks, we find a significant portion of GRB target genes, but few bystander genes [37].

The transcriptional co-activator p300 does not bind to DNA itself, but associates with enhancer-binding transcription factors on active enhancers. It has been shown to be associated with enhancers in cell lines [41, 42] and in vivo in mouse embryonic forebrain, midbrain and limb tissues [43]. The latter study detected between 561 and 2543 sites of p300 occupancy in each of the tissues, with most peaks located >10 kb from transcript start sites and about 90% overlapping sequences that are under evolutionary constraint. About 400 sites were marked by p300 in at least two tissues, suggesting that a fraction of enhancers were active in multiple tissues. The authors cloned the orthologous human sequence for 86 p300 occupied segments and obtained reproducible reporter gene expression in transgenic mice in at least one of the three tissues for 87% of the tested elements, improving on the typical success rate of enhancer prediction based on comparative genomics approaches (50–70% [3, 26]).

In a recent effort towards linking the epigenetic state of enhancers to expression programs, Heintzman et al. [41] showed that enhancers marked by p300 binding in ENCODE regions were enriched for H3K4me1 and frequently marked by acetylation of H3K27, DNaseI hypersensitivity, and binding of transcription factors and co-factors in patterns that were often cell-line specific and correlated well with strong evolutionary conservation.

The genome-wide binding and epigenetic modification data is about to transform the way we detect regulatory elements and select them for further experimental validation, by providing robust experimental data that will reduce the number of false positive regions and vastly reduce the sequence space in which to look for functional elements.
Currently, the UCSC Genome Browser provides tracks with ChIP-chip and ChIP-seq binding data from a number of experiments in human and mouse. Modencode (http://www.modencode.org [44]) strives to serve as a central repository for annotation of functional elements, including chromatin marks, in the D. melanogaster and C. elegans genomes. A central repository for other processed ChIP-based data is currently lacking, and they are currently not well integrated with those resources where they would potentially be most useful but must usually be extracted from literature or downloaded from author websites for upload as custom tracks to one of the genome browsers.

DATABASES OF FUNCTIONALLY ANNOTATED LONG-RANGE REGULATORY ELEMENTS

Two of the databases mentioned above, VISTA Enhancer Browser and CONDOR, contain information on the ability of conserved elements to drive expression of a reporter gene in model organisms. The annotation includes information on the developmental stage and anatomical structure where expression was recorded, which genomic fragment was tested, the primers used to clone that fragment, information on nearby genes, and links to genome browsers for further exploration of the genomic region.

The embryonic tissues displaying reporter gene expression indicate where the conserved element is active. To account for experimental variation, such as position effects (related to where in the genome the reporter construct happens to integrate), expression in a particular anatomical region is only recorded as positive in VISTA Enhancer Browser if reproduced in three independent mouse transgenic lines. The experimental data in CONDOR were produced using a similar strategy in the zebrafish model system, with the difference that CONDOR relies on superimposed transient expression locations (in 25 or more embryos), rather than stable transgenic lines.

The number of tested elements in VISTA Enhancer Browser was nearly 1100 at the time of writing. However, even combined, the two databases lack information on the great majority of HCNEs (hundreds of thousands in the vertebrate genomes). Moreover, the tested elements have only been examined at one or a few developmental time points. Therefore, lack of information or a negative result for a particular element does not imply that it is not a functional enhancer.

Other databases collect information on regulatory elements from the literature. For Drosophila, REDfly (http://redfly.ccr.buffalo.edu/; [45]) currently contains a collection of 727 empirically validated cis-regulatory modules (enhancers), their constituent TFBS, expression pattern and associated target genes. The ORegAnno database (http://www.oreganno.org; [46]), covers more species and includes several large-scale ChIP-seq and ChIP-chip studies, as well as data from more specialized databases such as REDfly. For some assemblies, UCSC and Ancora have tracks showing the elements from ORegAnno.

DATABASES OF IN SITU GENE EXPRESSION: ANNOTATING ENHANCERS THAT DRIVE REPORTER EXPRESSION

After obtaining the expression pattern driven by a tested enhancer, the target gene may be identified by comparison of the enhancer expression to the expression patterns of the genes in the (syntenic) genomic neighbourhood. The enhancer expression is expected to recapitulate (parts of) the specific expression pattern of the target gene. Thus, genes that are ubiquitously expressed and enhancers that, when tested in isolation, drive unspecific expression, cannot be analysed by this approach. A growing catalogue of online databases of in situ gene expression provide data in the form of images and/or expression annotation. The utility of these resources is somewhat hampered by lack of annotation and inconsistent use of anatomical ontologies, but they nevertheless provide an important means towards identifying target genes.

VisiGene (http://genome.ucsc.edu/) is an anatomical expression browser that currently shows expression from several repositories for mouse and one for Xenopus. The Allen Brain Atlas (http://www.brain-map.org/, [47]) contains in situ images and gene expression data for human cortex and mouse brain and spinal cord. ZFIN (http://zfin.org, [28]) is the central repository for zebrafish in situ expression data, and provides in situ images and gene expression annotation by developmental stage and anatomical location. The Berkeley Drosophila Genome Project (http://www.fruitfly.org) has produced a comprehensive in situ database for
D. melanogaster [48]. Flybase (http://flybase.org, [49]) provides additional expression data for Drosophila transcripts, proteins and reporter constructs from a wide range of sources as part of their gene reports. WormBase (http://www.wormbase.org, [50]) carries C. elegans gene expression patterns and expression profiles, and Xenopus in situ data are available in Xenbase (http://www.xenbase.org/, [51]).

POSSIBLE WORKFLOWS
A number of different tasks can be accomplished using a combination of the above resources. Some of the more common applications are given below.

Finding the target gene
Direct the Ancora browser to your region of interest or, using the HCNE density track, identify peaks of HCNE density and zoom in on a region. By way of example, here is how to use comparative genomics to identify the plausible target gene in that region. In some loci, e.g. human SOX3, there is only one gene surrounded by a gene desert and an array of HCNEs, making target gene identification straightforward. In other cases, e.g. the LHX1 locus (Figure 1), multiple genes are encompassed by the human HCNE array and human:teleost synteny block. Surveying which gene classes the genes in the GRB belong to reveals that LHX1 is a transcriptional regulator containing a homeobox domain and is involved in control of differentiation and development, a gene class commonly regulated by long-range elements. An inspection of the CpG island track reveals that the LHX1 gene, but none of the other genes in the GRB, is surrounded by multiple CpG islands - a hallmark of target genes. Additionally, in this deeply conserved locus, clicking on the synteny tracks to inspect human:zebrafish synteny reveals that lhx1a, one of the two zebrafish orthologs, is found on zebrafish chromosome 15 surrounded by the complete set of HCNEs, with none of the other genes present in that locus, further supporting the conclusion that those HCNEs are regulatory elements governing the expression of LHX1.

Identifying regulatory elements for a target gene
The HCNE density and synteny information in Ancora can be used to determine the extent of the genomic region in which the regulatory elements reside (Figure 1). Due to varying evolutionary distances between species and varying degrees of sequence turnover in different genomes and genomic regions, no hard rules can be given on what thresholds should be employed for comprehensive detection of HCNEs. Typical thresholds are in the range of 60–100% sequence identity over at least 30 bp. If the pre-computed elements are not sufficient for identification of individual conserved elements, the user may first identify the region of interest in Ancora on the basis of HCNE density and synteny, and then switch over to the VISTA, UCSC or Ensembl browser where individual elements may be selected by conservation, transcription factor binding site content, ChIP-seq data or some other criterion.

Examine if a non-coding SNP may be associated with changes in GRB regulation
Currently, Ancora does not contain any SNP tracks. To view a SNP in Ancora (i) search for it in the Ensembl or UCSC genome browser using its rs identifier (e.g. rs35989819); copy and paste its genome coordinates into the Ancora search window; (ii) optionally obtain the linkage disequilibrium (LD) block that encompasses the SNP, e.g. using HapMap data, to identify the region that would most likely harbor a causative SNP; (iii) Zoom out. In case of human or mouse SNPs, you will sometimes have to zoom out to a 1 Mb region or larger to see the entire possible GRB region; (iv) if the SNP or LD block appears to fall within a GRB region, find the most likely target gene for long-range regulatory elements in the region as above. (v) Examine the regulatory potential of HCNEs that overlap the SNP or LD block as above. To find SNPs within HCNEs, intersect your conserved regions of interest with SNPs using the UCSC Table Browser or Galaxy resources described below, or use the ECR browser, which provides lists of public SNPs for conserved regions alongside coloured bars indicating what proportion of the conserved sequence is coding and non-coding.

As can be seen, the current state of resources still calls for switching between resources that are not fully integrated, and some prior knowledge to decide on the genes that are most likely targeted by long-range regulatory elements.
BEYOND WEB BROWSERS AND DATABASES

When medium or high-throughput approaches are desired, genome browser methods fall short of providing enough functionality for simultaneous selection of a large number of potential elements to test or analyse. There are several tools available that can be used for this purpose.

Intermediate: web-based table manipulation

UCSC Table Browser (http://genome.ucsc.edu, [52]) is a web interface to tabular data underlying the UCSC genome browser. It retrieves genomic features corresponding to any coordinate-based track from the UCSC browser, with functions for filtering by regions, annotation or quantitative parameters. Its output can be sent to Galaxy.

Galaxy (http://galaxy.psu.edu/, [53]) is a powerful system for large-scale sequence-feature manipulation, probably the most powerful short of actual programming-based approaches (see below). Getting acquainted with it is highly recommended for anyone who works with large amounts of genomic coordinate-based data, although it can be useful for manipulating other types of data as well.

Advanced: programmatic approach

Most of the described data is processable with programmatic tools. For true power in genome-wide analysis, it is advised that ambitious beginning students in functional genomics acquire fundamental skills in one scripting language (Perl or Python), and in R with Bioconductor. The amount of available data and its power for new hypothesis generation are increasingly pushing such proficiency towards the status of a part of the standard skill set of future biologists. For readers new to table-based or programmatic data access, two recent tutorials on querying genome databases are a good place to start [54, 55].

PERSPECTIVE

Comparative genomics is a powerful tool in identifying genomic regions where long-range cis-regulatory elements are abundant, and to identify the genomic ranges around target genes at which long-range elements may be found. As we extend and improve the catalogue of assembled genomes, we will increasingly be able to use sequence and synteny conservation to identify long-range regulatory elements and the genes that are targets of such regulation across a wide range of species. Following identification of larger sets of target genes, we begin to probe the underlying mechanisms for how this regulation works and how it evolves. While GRBs regulate master developmental genes, long-range regulation also target a large number of other genes where the number and conservation of regulatory elements and their maximum distances from the genes are nowhere nearly so extreme. The findings from enhancer assays and epigenetics screens in a limited number of cell types indicate that there is a vast number of enhancers in metazoan genomes that are used to drive specific gene expression at different developmental stages and in different cell types at different conditions. The tools presented here are currently the best entry points for the further study of these elements. Further integration, more data, and more powerful visualization analysis tools are expected in the very near future.

Key Points

- Regions containing long-range regulatory elements driving the expression of developmental regulatory genes (genomic regulatory blocks or GRBs) can be identified by their high density of HCNEs.
- The main questions about GRBs that can be answered computationally are (i) what is the target gene in the region, and (ii) what is the extent of the regulatory elements in the region. A number of web-based tools available to answer these questions are described and compared.
- Further hypotheses about the developmental genes and their regulation can be generated by using additional epigenomic and expression data. We describe several use cases that include combined use of multiple tools and data sources.

Acknowledgements

We thank Thomas S. Becker for illuminating discussions and critical reading of the article.

References


