Genome-wide enhancer – gene regulatory maps in two vertebrate genomes

Yves Clément*, Patrick Torbey, Pascale Gilardi-Hebenstreit & Hugues Roest Crollius*

Ecole Normale Supérieure, PSL Research University, CNRS, Inserm, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), F-75005 Paris, France

Corresponding authors:
YC: yclement@biologie.ens.fr
HRC: hrc@ens.fr

Abstract

Gene regulation by enhancer sequences controls the spatiotemporal expression of target genes, often over long genomic distances. Current methods used to identify these interactions in the human genome require complex experimental setups and are restricted to specific cell types. Here we use PEGASUS, an approach that captures evolutionary signals of synteny conservation to predict such interaction potentially in any biological context. We apply it to the human and zebrafish genomes and exploit the 1.3 million human and 55,000 zebrafish enhancers that we predicted to uncover fundamental principles of gene enhancer function in vertebrates. We show that the number of enhancers linked to a gene positively correlates with expression breadth in space and time and that the enhancer-target distance is evolutionarily neutral. We uncover almost 2000 regulatory interactions ancestral to vertebrates, which are strongly enriched in core developmental processes. Using PEGASUS and its evolutionary view of enhancer-gene interactions, we provide a highly complementary resource to functional assays which uncovers key principles of enhancer biology.

Introduction

Enhancers are short DNA sequences that bind transcription factor and contact promoters in *cis* to activate or repress the transcription of protein coding genes into messenger RNA[1]. This control – or regulation – of gene expression by enhancers ensures the fine tuning of protein abundance in cells, in the course of their differentiation during development but also in response to various stimuli in specialised cell types[1-3]. Understanding how enhancers regulate gene expression is therefore essential to understand how organisms develop and function throughout life. In fact, disruption of enhancer function has been shown to lead to abnormal gene expression and thus to diseases[4-8]. In addition, the majority of variants identified in Genome Wide Association Studies (GWAS) are found outside coding sequences[9,10]. Together with the observation that many patients remain undiagnosed after genome sequencing because no plausible coding variant can be incriminated[11], these results underscore the importance of identifying enhancers and their target genes to better understand genome biology.

Numerous methods have been developed to identify enhancers. Some are based on the analysis of the conservation of non-coding sequences[12-15]. Others analyse the binding of enhancer-associated proteins on the genome[16,17], or identify chromatin signatures (e.g. histone modifications[18-20] or open chromatin[21,22]). However, identifying the target genes of a given enhancer remains difficult because enhancers can regulate genes located at long distances, typically several hundreds of kilobases (kb) away[23,24]. Methods that have been developed to address this problem fall in three broad categories. The first uses chromosomal conformation capture techniques to identify physical interaction between two loci in the genome[25-30], but it is not clear which of these interactions are linked to a regulatory role. The second measures the correlation of transcription activity between non-coding sequences and nearby genes[31,32], assuming the two are signatures of a coordinated regulatory function. Finally, single nucleotide variants (SNVs) can be associated to significant differences in gene expression, thus qualifying as expression QTL (eQTL) that presumably reside within or close to enhancers[33]. These methods all rely on a specific experimental set up, i.e. a particular cell line or tissue upon which the experiment can be carried out, therefore investigating regulatory interactions in a specific but narrow biological context. We previously developed PEGASUS (Predicting Enhancer Gene ASsociations Using Synteny), a method to identify the target genes of predicted enhancers using an evolutionary signal instead of an experimental assay[34], therefore potentially identifying all functional interaction

that are under negative selection and hence conserved. This is of particular interest for regulatory functions important during human development, where experimental assays are limited. We tested this method on the human X chromosome followed by experimental validations of more than 1,000 predicted interactions using transgenic assays[34].

Here, we applied this method on the entire human and zebrafish genomes in two independent analyses. We predict the interactions between 1,311,643 human enhancers and 18,339 target genes, and between 55,515 zebrafish enhancers and 17,363 target genes. By comparing human and zebrafish predictions, we outline a set of genes with conserved cis-regulation in vertebrates enriched in brain and development functions. We find evidence for a direct link between the expression breadth of human genes and the number of enhancers predicted to target them, explaining at least in part the broader expression patterns of genes with promoters overlapping CpG islands. We find that the average distance separating enhancer and their target genes scales with genome size, demonstrating that little selective pressure acts to preserve this distance. In addition, highly conserved enhancers are more likely to reside within the same Topologically Associating Domain (TAD) as their target gene, irrespective of the distance separating them. This evolutionary signal provides a basis to explain the observed conservation of TADs in vertebrates[35,36] and sheds light on the complex relationship between 3D genome organisation and the regulation of gene transcription.

Results

Identification of enhancers & target genes in two vertebrate genomes

We applied the PEGASUS method[34] to identify predicted enhancers and their most likely target genes in the human and zebrafish genomes. We first analysed the UCSC 100-way multiple genome alignment restricted to Sarcopterygii (all vertebrates except fish but including Coelacanth) to identify human conserved non-coding elements (CNEs). We then computed an evolutionary linkage score between each CNE and neighbouring genes within a 1 Mb radius, thus identifying the linkage that is the most conserved. This resulted in predicting 1,311,643 CNEs linked to 18,339 human genes. We replicated the analysis in zebrafish using a multiple alignment of 7 teleost fish genomes, leading to the identification of 55,515 CNEs linked to 17,363 genes. Of note, the zebrafish was compared to 6 genomes and shares a 300 million years old ancestor with the closest species of this set, while in contrast human was compared to 35 genomes, including many placental mammals who share a ~90 million years old common ancestor (Supplementary Figure 1). By reducing the phylogenetic sampling used to analysing the human genome to a situation equivalent to that of zebrafish, we show that this contrast is sufficient to explain the vastly different numbers of CNEs identified in zebrafish (Supplementary material). Therefore, the number of zebrafish CNEs identified here are likely to be substantially underestimated. CNEs are linked to a median number of 3 genes in human and 4 genes in zebrafish, while genes are linked to a median number of 301 enhancers in human and 5 enhancers in zebrafish. These CNEs are short, their size ranging from 10 bp to 1,709 bp in human (with a median of 27 bp), and from 10 bp to 1,628 bp in zebrafish (with a median of 90 bp; Table 1). CNEs are relatively close to their linked genes: the median distance to TSS was 353,488 bp in human and 288,643 bp in zebrafish (Supplementary Figure 2), with 29% of enhancer - TSS distances in human and 37% in zebrafish smaller than 200 kbp.

Because PEGASUS can predict CNE - gene interactions that skip over unrelated genes, we defined three categories of CNEs: intronic (located on an intron of their linked gene), flanking (intergenic CNE flanking their linked gene, and jumping (CNEs skipping over intervening genes to connect to their linked genes). We found that 20% of predicted

enhancers are intronic, 53% flanking and 27% jumping in human, while in zebrafish, 15%, 36% and 49% of predicted enhancers are respectively intronic, flanking and jumping. Interestingly 34% of intronic human CNEs (37% in zebrafish) are linked to a gene that is not their "host" gene. These results suggest that choosing the nearest neighbour as potential target gene, as if often done short of a better alternative, is likely to fail in more than 27% of cases (49% in zebrafish).

	hg19	***danRer7***
# of enhancers	1 311 643	55 515
# of target genes	18 339	17 363
median length (bp)	27	90
median distance to TSS (bp)	356 366	290 748
median genes per enhancer	3	4
median enhancer per gene	301	5
% of intronic / flanking / jumping enhancers	19.94 / 53.09 / 26.97	15.04 / 35.57 / 49.39

Table 1: Statistics about predicted enhancers in human and zebrafish.

From CNEs to enhancers

PEGASUS CNEs are solely defined by sequence conservation, from weak to strong, yet all CNEs are assigned a linkage score as long as a gene lies within a 1 Mb radius. However, not all CNEs are regulatory enhancers and we wished to confirm that the linkage score is correlated with evidence for a regulatory role in the human and zebrafish genomes, as noted previously for human chromosome X[34]. CNE linkage scores are strongly correlated in both species with chromatin marks (H3K27ac, H3K27me3, H3K4me1 and H3K4me3) in human embryonic stem cells[20] or in various developmental stages in zebrafish[37] (Figure 1 A & B, Supplementary Tables 1 & 2). Interestingly, the two marks associated with active enhancer function (H3K4me1 and K3K27ac[38]) show the strongest overlap (Z-score = 85.33 and 21.55 for human, 160.51 and 144.57 for zebrafish respectively). A strong correlation between CNE linkage scores and open chromatin marks (ATAC-seq in zebrafish and DNase1 marks in human) further supports the regulatory role of linked CNEs (Supplementary Figure 3). In addition, CNEs for which no target gene could be identified have almost no experimental evidence for a regulatory function, showing that the PEGASUS linkage score also recognises non-functional CNEs (Figure 1 and Supplementary Figure 3). Finally, the overlap between CNEs and experimentally verified enhancers in human[31,39] and zebrafish[40] follows the same notable enrichment profile as a function of the linkage score (Figure 1 C & D). Together these results show that human and zebrafish CNEs identified here are strongly enriched in functional enhancers.

PEGASUS target gene predictions overlap experimentally defined regulatory interactions

We sought to measure the agreement between PEGASUS target gene predictions and those from tissue or cell-type dependent experimental approaches, restricting our analysis to the 30% of CNEs linked to a single gene. While 28% of our target gene predictions agree with Capture Hi-C interactions[27,30], the level of agreement depends on cell type: from 22% for Embryonic Stem Cells (ESC) to 29% for GM12878 (a lymphoblastic cell line). The level of agreement with eQTLs-based predictions[33], is higher, with a 57% overlap. Interestingly, in all dataset that we analysed, the agreement decreased with distance to the TSS (Supplementary Figures 4), even in the range 0-500 kb where the linkage score is relatively constant (Supplementary Figure 5). This could be explained by the observation that longer distance CNEs are active in fewer tissues (Supplementary Figure 6), limiting the possibility of overlap

with dataset obtained on a limited number of cell or tissue types. The level of agreement between PEGASUS predictions and methods based on functional assays is within the same range as when the latter are compared to each other (between 10 and 48% of eQTLs-based predictions agreed with Capture Hi-C predictions, depending on the cell type), and here also, the overlap decreases with the distance to the TSS (Supplementary Figure 6). We therefore conclude that PEGASUS predictions enhancer-TSS predictions are consistent with results from genome wide functional assays.

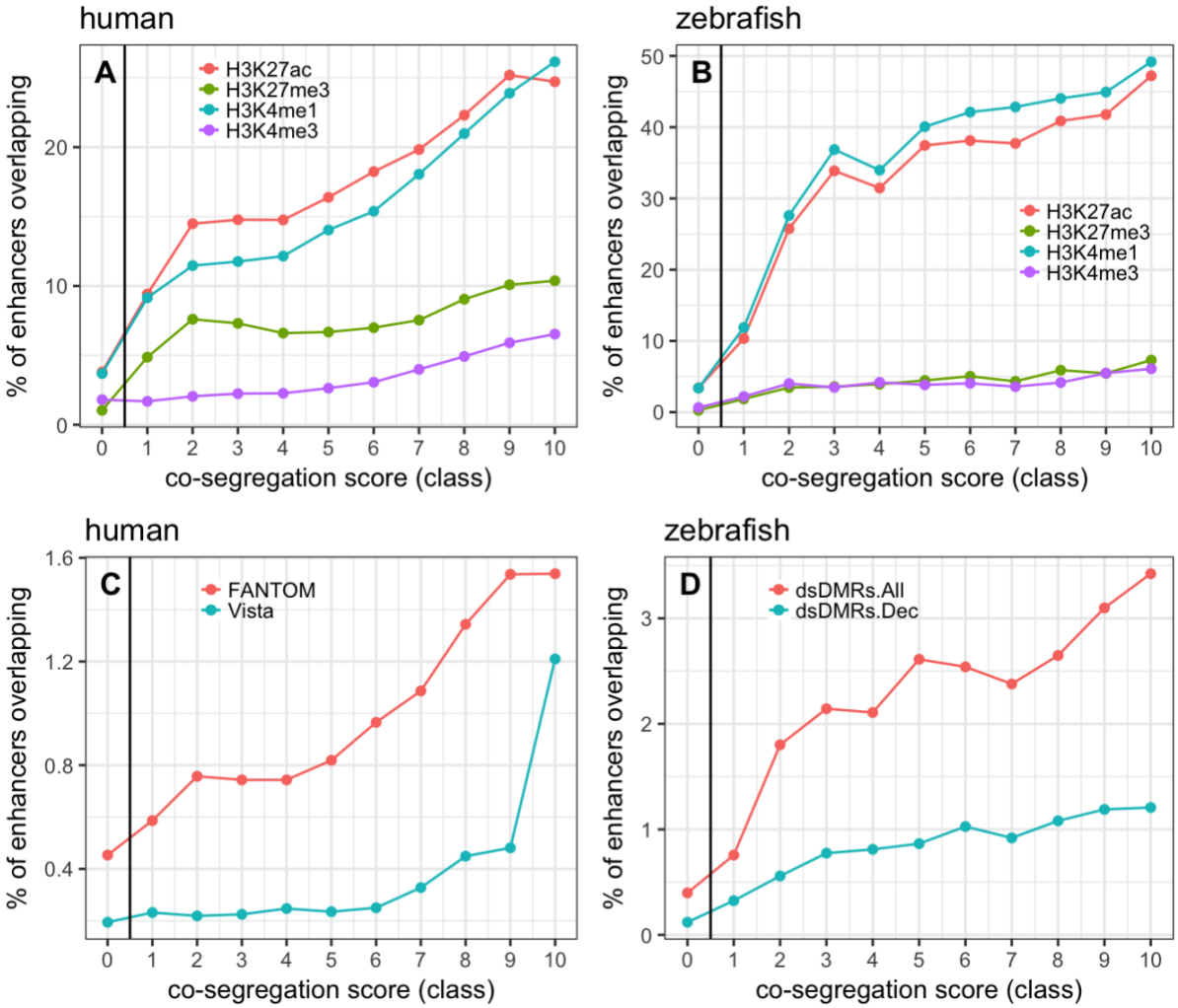

Figure 1: Overlap between functional annotations and predicted enhancers.
Percentage of predicted enhancers overlapping chromatin modifications (top line) or enhancer predictions (bottom line) in human (left column) and zebrafish (right column) in classes of co-segregation score. Predicted enhancers were divided into 10 classes of identical size according to their co-segregation scores. Class 0 represents enhancers with no predicted target gene. The top line represents overlap with ChIP-seq peaks of histone modifications in human embryonic stem cells[20] or in various developmental stages in zebrafish[37]. The bottom line represents overlap with enhancer predictions from FANTOM[31] or Vista[39] in human, and from differentially methylated regions during development in zebrafish[40].

Long-range regulatory interactions within TADs

Topologically Associating Domains (TADs[35]) have been shown to coincide well with the regulatory landscape governing gene expression[6,41]. Here, 57% and 66% of CNEs linked to a single gene indeed reside within a TAD in hESCs and IMR90 cells respectively, compared to

an average of 34% and 40% respectively when we shuffle TAD intervals. CNEs are linked to their target gene with a higher score when inside the same TAD (mean scores, 0.72 inside vs 0.67 outside for hESCs, 0.71 vs 0.68 for IMR90, Figure 2), overlapped more with functional marks (13.9% vs 9.7% for H3K4me1, 14.9% vs 11.1% for H3K27ac in hESCs) and were also closer to each other (median distance to TSS, 332 kb inside vs 522 kb outside for hESCs, Figure 2). Finally, we see a striking link between predicted enhancer conservation and their localisation within a TAD. First, enhancers-target gene pairs located within the same TAD have more often orthologous enhancer-targets conserved in zebrafish than other pairs (0.23% vs. 0.18%). More importantly, we see a positive link between the conservation depth of enhancers and the co-localisation of enhancers and target genes within the same TAD (Supplementary Figure 7). The association between TAD co-localisation and both distance to TSS and conservation depth cannot be explained by chance alone (Supplementary Figure 7). In line with previous observations[42], our results therefore support the existence of a link between the control of gene expression and 3D chromatin organisation.

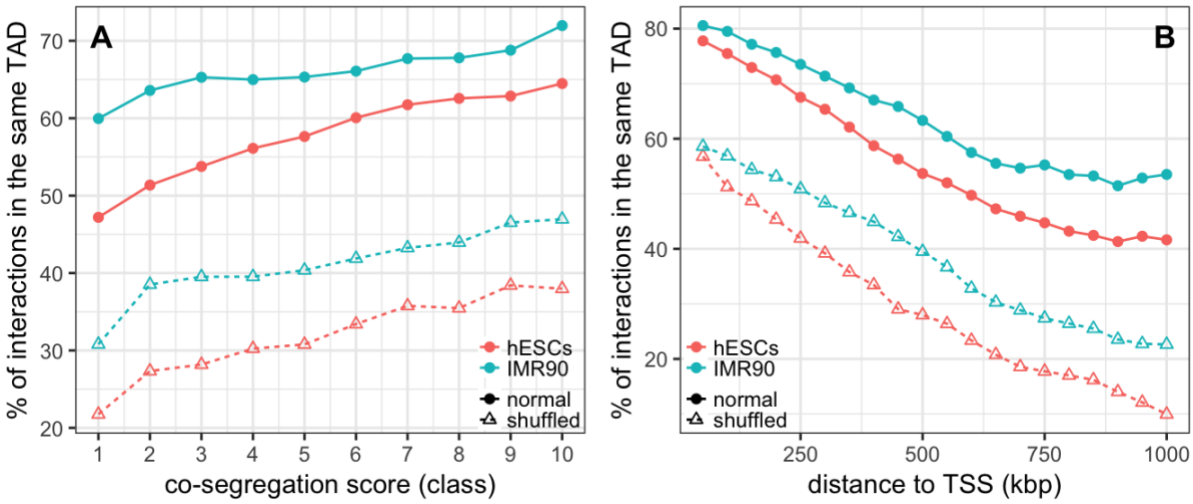

Figure 2: Overlap between topologically associating domains and regulatory interactions.

Percentage of predicted enhancer – target gene interactions located in the same TAD as a function of distance (left panel) or number of species in which the predicted enhancer is conserved (right panel). TADs were obtained from two cell types (IMR90 and hESCs[35]).

Function of regulatory interactions conserved in vertebrates

The human and zebrafish CNE-gene maps described here were built independently using distinct sets of genome, allowing us to use their intersection as increased evidence for functional and highly conserved enhancer - target gene interactions. In addition, these interactions were probably inherited from their last common Euteleostomi ancestor. We identified 1,986 human CNEs and 1,949 zebrafish CNEs conserved between human and zebrafish (methods) and linked to 567 human genes and 608 zebrafish orthologous genes respectively. Gene Ontology term analysis (methods) shows that these ancestral regulatory circuits are overwhelmingly enriched in neuronal, skeleton, muscle, vascular and renal development (Table 2). To illustrate this, we used the CRISPR-Cas9 technology to delete a 236 bp zebrafish CNE predicted to target the irx1b gene. Its orthologous CNE targets IRX1 in the human genome. This gene plays multiple roles during pattern formation of vertebrate embryos[43], and we expect its expression pattern to be tightly regulated by a complex array of enhancers. The deletion of the CNE greatly decreases the expression of the endogenous gene in several structures of the zebrafish embryo (Figure 3) establishing it as a *bona fide*

developmental enhancer. The strong enrichment in core developmental functions observed with orthologous PEGASUS predictions is consistent with earlier observations, as enhancers identified through sequence conservation are often found to be active during development, especially in the nervous system[12-14,44-46].

hg19			***danRer7***		
GO term	*fold enrichment*	*p-value*	*GO term*	*fold enrichment*	*p-value*
ventral spinal cord interneuron differentiation	14.02	4.91E-02	potassium ion import	11.28	1.30E-02
positive regulation of cardiac muscle cell proliferation	11.38	3.18E-02	central nervous system neuron differentiation	6.99	2.73E-04
central nervous system projection neuron axonogenesis	11.38	3.18E-02	embryonic cranial skeleton morphogenesis	5.79	1.16E-04
proximal/distal pattern formation	9.35	2.73E-02	cell fate commitment	4.88	3.29E-02
cell fate determination	8.90	2.54E-03	positive regulation of transcription from RNA polymerase II promoter	4.21	1.54E-02
forebrain neuron differentiation	6.80	2.77E-02	regulation of neurogenesis	4.03	6.57E-03
neuron fate commitment	6.50	4.67E-03	camera-type eye development	3.43	1.67E-02
neural crest cell differentiation	6.14	2.37E-02	tissue morphogenesis	3.08	8.67E-04
renal tubule morphogenesis	6.14	2.37E-02	transcription, DNA-templated	2.97	1.54E-07
nephron tubule development	5.90	1.26E-02	tube development	2.92	2.43E-02

Table 2: Top 10 overrepresented GO terms in human – zebrafish orthologous genes with conserved regulation

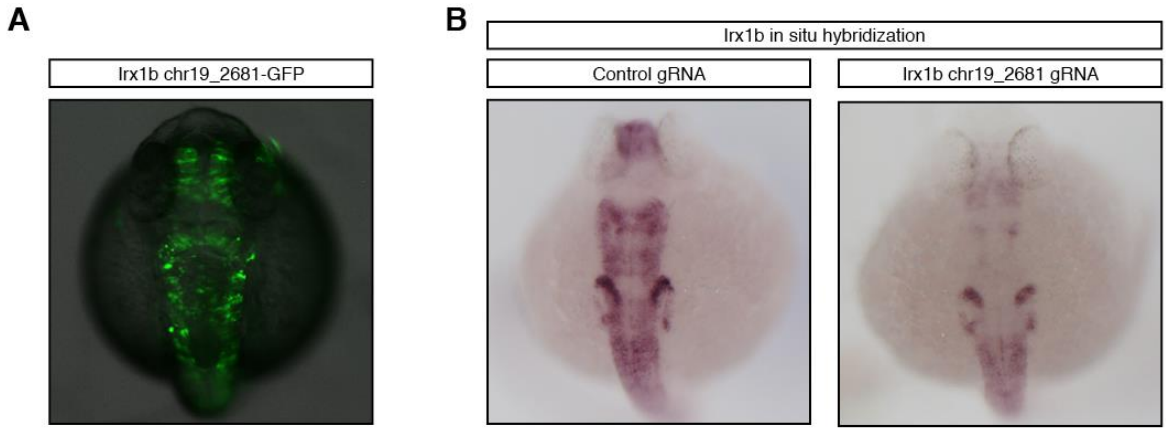

Figure 3: Example of in vivo validation of predicted enhancers
(A) 24h old F0 zebrafish embryos injected with a Tol2 transposon containing the predicted *Irx1b* CNE positioned 5' of the gata2 minimal promoter driving green fluorescent protein (GFP) expression. (B) In situ hybridization for *Irx1b* mRNA performed on 24h old embryos injected with a mix of 3 CRISPR/Cas9 ribonucleoprotein complexes targeted against a random intergenic locus (control), or the predicted *Irx1b* enhancer. The CNE activity profile overlaps with *Irx1b*'s expression profile, which is comprised of the acousticovestibual ganglia, tectum, tegmentum, cerebellum, the hindbrain, the spinal chord and the lateral floor plate, but not in the mid-hindbrain boundary. *Irx1b*'s expression level is greatly decreased in all these structures when the CRSPR/Cas9 complex is targeted to the predicted enhancer compared to the control, establishing it as a bona fide *Irx1b* enhancer.

CNE – gene distances scale with genome size

The "action range" of enhancers is known to encompass a wide span, from within the target gene itself to more than 1 Mb away. It is therefore reasonable to postulate that little constraints exist on maintaining a specific genomic distance between enhancers and their target genes, especially as they are thought to physically contact each other through DNA looping in the nucleus. However, to our knowledge, this has never been formally tested at a

genomic scale. The genomes of human and zebrafish are very different in size, 3.1 Gb and 1.5 Gb respectively, allowing us to examine this question using the ~ 2000 orthologous human-zebrafish CNE - gene linkage: do their respective genomic span scales with genome size or instead remain conserved to some degree? We used the sizes of orthologous introns to estimate the relative neutral evolution of genomic distances, since introns are thought to be under negligible size constraint above a small minimal threshold. Results show that distances between orthologous CNEs and their orthologous target genes scale precisely with intron size (Figure 4). There is therefore no evidence for a constraint on CNE – gene distance, suggesting that 3D folding principles governing enhancer – promoter contacts are not sensitive to this distance.

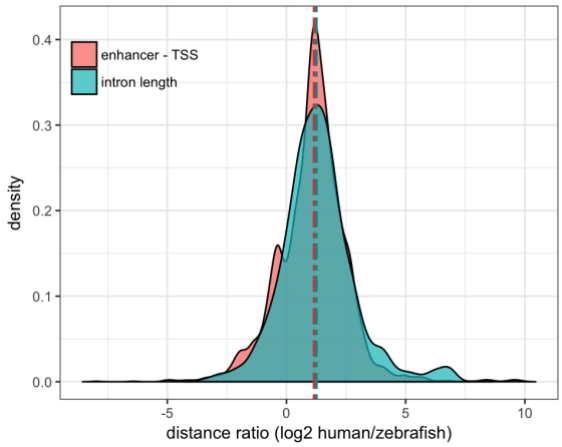

Figure 4: Comparison of enhancer – target genes interaction distances.
Pairwise comparison of interactions distances between human – zebrafish orthologous genes with conserved predicted enhancers, and of intron length in the same set of genes. Interaction distances are computed as enhancer-TSS distances. Comparisons are computed as log2 values of human/zebrafish ratios, with each pair of orthologous genes being studied independently.

Regulation complexity is linked to expression breadth

Enhancers are largely responsible for the spatio-temporal specificity of their target gene's expression (e.g. SOX2[47], SOX9[48] or HOX genes[49]). It naturally follows that the complexity of a gene's expression pattern might be correlated with the number of enhancers that regulates its expression, although this has never been demonstrated. We used our predicted enhancers to investigate this question. We observe a positive link between the number of enhancers targeting a gene and the number of adult tissues where this gene is expressed ($\rho = 0.23$, p-value $< 10^{-15}$, Figure 5). Genes cover a broad range of tissue specificities, from 'housekeeping genes' required for generic cellular functions and expressed in most tissues to tightly regulated developmental genes sometimes expressed in just a few cells in a short window of time. The presence of a high frequency of CpG dinucleotides in the promoter of genes (so called CpG islands) is a useful quantitative marker of broad expression[50-53]. We split our target genes between those which TSS overlap a CpG island (referred to as CpG genes) and other genes (referred to as non-CpG genes). First, we confirm that CpG genes are more broadly expressed than non-CpG genes (one-sided Kolmogorov–Smirnov test p-value $< 10^{-15}$, Supplementary Figure 8). Second, we observe that CpG genes are targeted by a higher number of predicted enhancers than non-CpG genes (365 vs 314 on average, Kolmogorov-Smirnov test p-value $< 10^{-15}$, Supplementary Figure 8), thus providing an immediate mechanism for their broader expression. The link between predicted enhancer number and expression breadth is stronger for non-CpG genes than for CpG genes ($\rho = 0.29$ & 0.16, respectively, p-values $< 10^{-15}$). But this is explained by the narrower (but higher)

range of tissue number where CpG genes are expressed (Figure 5). Indeed, reducing the range of expression for non-CpG genes to a level similar to CpG genes decreases correlations to similar levels as CpG genes (Supplementary Figures 9). Our results show that expression patterns are likely to be driven by regulation complexity, and that the effect is stronger for non CpG genes. We asked if, in addition to expression complexity, a similar influence of predicted enhancers could be observed for expression levels of human genes. However, the number of predicted enhancers only weakly correlates with expression levels in 17 out of 21 tissues (p-value <= 0.05, spearman's ρ from -0.08 to 0.15). Regulation complexity does not seem to be correlated with expression levels in human tissues.

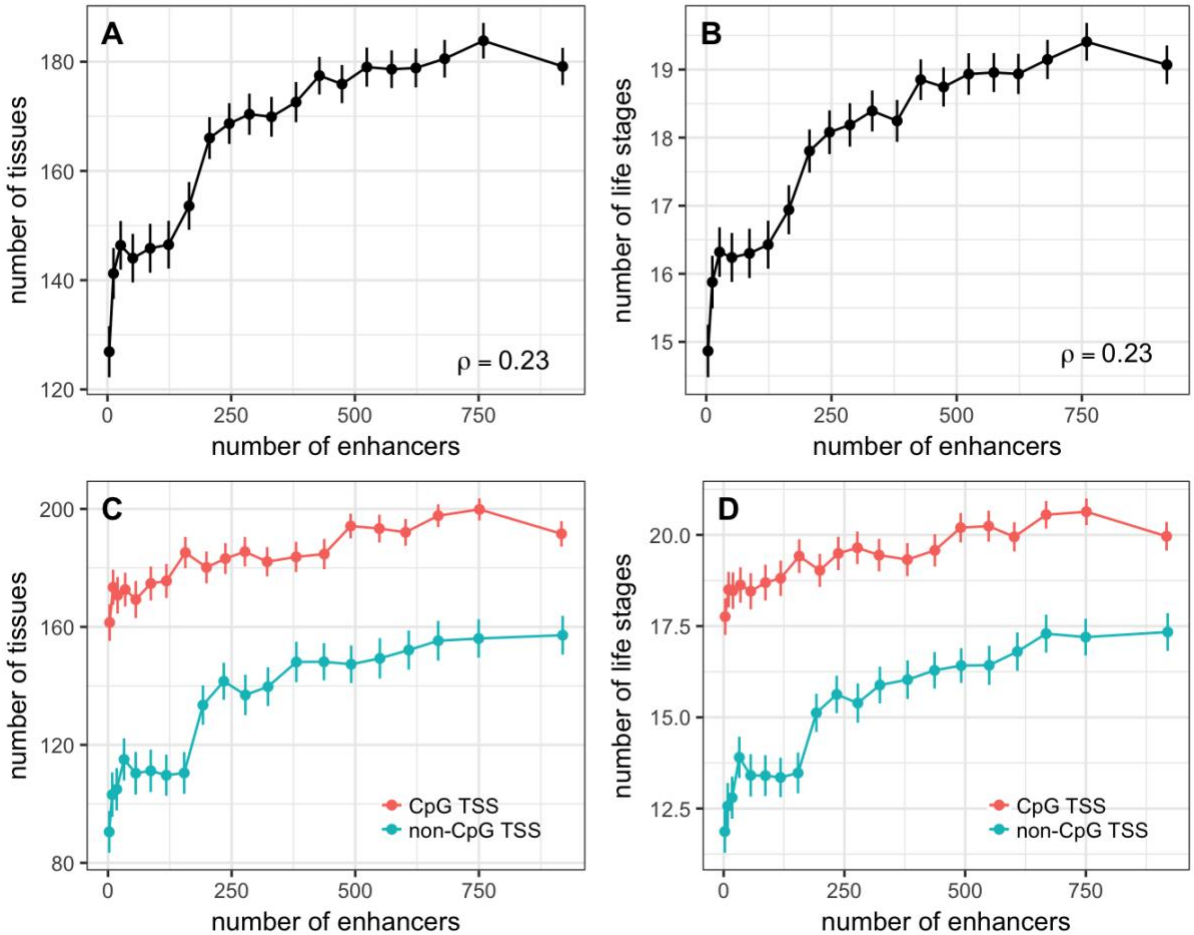

Figure 5: Link between regulation complexity and expression breadth

Link between the number of predicted enhancers targeting a gene and the number of tissues (left column) and life stages (right column) it is expressed in, for all genes (top line) or separating between CpG TSS genes and other (bottom line). Genes were divided in twenty classes of identical size based on their number of predicted enhancers. Points and vertical lines represent mean number of tissues or life stages with 95% confidence interval in each class. Correlation coefficients were computed on unbinned data. Genes were classified as CpG TSS or non-CpG TSS based on the overlap of their TSS with CpG islands.

Discussion

We applied the PEGASUS method to predict ~1,300,000 human and ~55,000 zebrafish enhancers targeting the majority of the genes in their respective genomes. After finding evidence for a regulatory role of these interactions, we used this dataset to make a number of fundamental observations related to enhancer function. We show that the number of enhancers that regulate a gene positively correlates with its breadth of expression, that the

distance between predicted enhancers and their target gene evolves neutrally, that chromatin organisation and regulation of gene expression are linked and that regulatory interaction ancestral to vertebrates concentrate on core functions necessary to build an organism (skeleton, muscles, neurons, vascular system).

The challenges of predicting long-distance regulatory interactions

Most methods currently employed to predict long-range regulatory interactions in the human genome rely on specific cell lines[26-31,33]. These methods usually differ in their approach and the cell types or tissues they study, which might limit expectations to observe overlap in their predictions, especially given that most enhancers are tissue-specific[29,54-56]. It is therefore interesting to observe that when comparing these methods with one another, the level of agreement between eQTLs-based predictions ranges from 10% to 48% (between 10% and 16% for ESCs and NECs cells, between 38% and 48% for CD34 and GM12878 cells). In contrast, PEGASUS is agnostic to cell-type or tissue context. The sole rationale underlying the predictions is that the interactions are functional, therefore under sufficient evolutionary conservation to be picked up by comparisons with other genomes. This property introduces its own bias, in that conserved interactions are more likely to relate to developmental processes[16,57], which are typically harder to identify in differentiated cell lines. We show that our predictions agree with functional assays in a manner consistent with the intersection between the latter studies, suggesting that an evolutionary-based approach can capture more than 50% of interactions observed in experimental assays. Remaining, non-overlapping interactions likely include false positives in either type of methods, which are difficult to quantify at present. Another source of discrepancy may be that PEGASUS misses cell-type specific enhancers that do not show enough sequence conservation to be identified by evolutionary linkage: given the rapid evolutionary turnover of enhancer regions during evolution[58-60], it is likely that a fraction of cell-type specific enhancers have had little time to leave detectable footprints of selection in a genome. We observed a general trend for predictions, whether experimental or evolutionary based, to overlap consistently less with increasing distance between the predicted enhancer and the TSS of the target gene. Our data suggests that this could be explained by long distance enhancers being more tissue-specific than short distance enhancers. This has important implications for regulatory interactions predictions.

No evidence for natural selection acting on enhancer – target gene distances

Our results suggest that there is no selective pressure acting on the enhancer – target gene interaction distance to ensure proper regulation. These results can easily be explained by how enhancer regulation is mediated through the 3D organisation of the genome. Metazoan genomes are organised in TADs (for topologically associating domains), large units of chromosomal interactions[61,62] that are mostly conserved between cell types and species[35,36,63]. Enhancer – target genes interactions occur mostly within TADs[64], via DNA looping[65]. Interestingly, it has been recently shown that the position of an enhancer relative to its target within a TAD has no effect on its regulatory potential[66]. These results are consistent with our finding that the relative position of an enhancer relative to its target gene is effectively neutral. As a result, the same forces that affect genome sizes will affect the interaction distances. A recent analysis of genomic regulatory blocks (or GRBs) in metazoans based on the analysis of clusters of conserved non-coding elements showed that a) these blocks correlated well with known TADs and b) their sizes seem to correlate well with genome size[67], providing further evidence that interaction distances between enhancers and target genes are under the same forces that affect genome size in metazoans.

Conclusions

This study provides a unique view of the conservation and evolution of enhancers in vertebrate genomes. Our results support the model where gene expression complexity is driven by regulatory region diversity and highlight the biological functions which regulation has been conserved since the vertebrate ancestor. Moreover, the PEGASUS method provides a robust tissue and life stage agnostic target gene prediction method that opens research possibilities in the study of gene regulation in a wide number of species.

Materials & Methods

Defining conserved non-coding elements and their most likely target genes

We used a previously published method to identify predicted enhancers and their most likely target genes[34]. This method first identifies predicted enhancers as conserved non-coding elements (or CNEs for short) in multiple alignments, and second identifies a predicted enhancer's most likely target gene as the gene in its vicinity with the most conserved synteny, through the computation of a co-segregation score measuring this conservation.

We identified CNEs in the human genome (*hg19* version) using the UCSC 100-way multiple alignments that we restricted to 35 Sarcopterygii species (a full list is available in Supplementary Table 3). CNEs had to be identified in at least 6 species (including human) with one non-primate species to be considered. We identified CNEs in the zebrafish genome (*danRer7* version) in custom-made multiple alignments that include 6 other Neopterygii species (a full list is available in Supplementary Table 4). Multiple alignments were built first by making pairwise alignments between zebrafish and other species using LastZ[68], then by using these alignments to build multiple alignments using Multiz[69]. Alignment parameters can be found in Supplementary Materials. In these multiple alignments, CNEs had to be identified in at least 3 species (including zebrafish) to be considered. We used a seed of 10 bp (with 90% conservation) and a maximum fraction of mismatches for a column to be considered as pseudo-conserved of 40% for both analyses.

Most likely target genes were identified with the same method in both genomes. We first identified all protein-coding genes in a 1Mb radius around CNEs in the reference genome and computed a co-segregation score for each gene as a measure of the conservation of synteny between a CNE and a gene. Considering only species in which the CNE has been identified, the score sums over all species and adds a term when the gene is identified, on the same chromosome and in the radius, and subtracts a term when the gene is either absent from the genome, on a different chromosome or outside the radius. Terms are normalized by pairwise rearrangements rates and genome sequencing coverage (see[34] for more details). For a given CNE, the most likely target gene is the gene with the highest co-segregation score. Adjacent CNEs targeting identical gene, present in the same species, having identical co-segregation scores and distant by less than 100bp were merged together. Predicted enhancers located at 100bp or less away from an exon were discarded.

Overlap with functional marks & enhancer predictions

We analysed the overlap of our predicted enhancers with functional marks and previous enhancer predictions. We first looked at the overlap with ChIP-seq peaks of histone modifications (namely H3K27ac, H3K4me1, H3K4me3 and H3K27me3) in embryonic stem cells in human[20] and across various developmental stages in zebrafish[37]. We also looked at the overlap with peaks of ATAC-seq in zebrafish[70]. We also looked at the overlap of our

375 predicted enhancers with other enhancer predictions. In human, we looked at enhancers from the FANTOM project[31] or from the Vista database[39]. In zebrafish we looked at enhancer predictions from differentially methylated regions[40].

Gene expression data

Gene expression values and calls were downloaded from the Bgee database[71]. For each gene,
380 we computed the number of adult tissues for human or developmental stage for zebrafish in which a gene is called as expressed. We filtered out terms that had daughter terms in our lists of tissues or stages. For human, we also retrieved expression levels in 21 tissues.

Comparing target gene predictions with chromosome conformation capture predictions

385 We compared our enhancer - target gene predictions in human with other predictions based on correlation of transcriptional activity between predicted enhancers and target genes[31], on eQTLs in various tissues[33] or on Capture Hi-C data in CD34 and GM12878 cells[27] and in embryonic stem cell (ESCs) and ESC-derived neurodectodermal cells (NECs)[30]. The latter contains interactions present in either cell types or in both simultaneously
390 (labelled "Both"). We compared predictions between our datasets and others firstly by identifying overlapping enhancers between our predictions and other datasets, then by identifying identical interactions predictions in these overlapping regions.

Defining orthologous enhancers & target genes between human & zebrafish

We downloaded human-zebrafish and zebrafish-human pairwise chain alignments from
395 UCSC. We defined orthologous predicted enhancers as human and zebrafish enhancers that overlapped by at least 10bp on either pairwise alignment. We next downloaded human-zebrafish orthologous genes from the Ensembl database (version 75) to identify orthologous enhancers targeting orthologous genes.

Because of the evolutionary distance between human and zebrafish, some orthologous
400 regions are difficult to align and are thus impossible to detect. To circumvent this problem, we used the spotted gar genome[72] to identify orthologous predicted enhancers. We downloaded human-spotted gar pairwise chain alignments and used our custom-made zebrafish-spotted gar pairwise chain alignments to respectively map human and zebrafish predicted enhancers on to the spotted gar genome. We considered human and zebrafish
405 enhancers as orthologous if they overlapped by at least 10bp on the spotted gar genome. We identified orthologous targets by looking at the orthologous genes set used above. Both direct and via spotted gar sets of orthologous enhancers – orthologous target genes were merged together.

Gene expression enrichment analysis

410 We performed GO terms enrichment analyses using the PantherDB website[73,74]. The website performs a Fisher's exact test with FDR multiple test corrections.

Distance to transcription start sites

We downloaded transcription start sites (or TSS) locations from the Ensembl database (version 75). For each gene, we considered only the transcript annotated in RefSeq giving
415 the longest protein. We computed for each enhancer - target gene the distance to the TSS as the shortest distance from enhancer boundary to the target's TSS.

Topologically associating domains

We downloaded topologically associating domains (or TADs) coordinates for two cell types, human embryonic stem cells (hESCs) and IMR90 fibroblasts[35]. We converted these coordinates from hg18 to hg19 using the liftOver webtool available at the UCSC genome browser[75]. For enhancers targeting only one gene, we computed for each cell type whether both an enhancer and its target gene were located in the same TAD. We also generated shuffled TAD set by shuffling the TADs coordinates along the human genome.

In vivo validations

Vector and cloning:

The predicted *Irx1b* enhancer was amplified from zebrafish genomic DNA using the following primers: CNE-Irx1b-Forward: 5'-TGAATGCTCATCCGGAACATCCACTGCTGCTCCCAAAG-3'; CNE-Irx1b-Reverse: 5'-GACCTGCAGACTGGCAGTTCCTCGCCAGAGCTCAG-3' and cloned into pZED plasmid[76] upstream of the minimal gata2 promoter/gfp reporter.

Zebrafish egg injections for transgenesis:

The Tol2 transposon/transposase method of transgenesis[77] was used with minor modifications. Two nanoliters containing 20 ng/µl of transposase mRNA and 30 ng/µl of phenol/chloroform purified pZED construct were injected in one-cell stage embryos.

In situ hybridization:

In situ hybridization were performed as described[78], using an *Irx1b* probe corresponding to exon 2.

Zebrafish egg injections for mutagenesis:

Three RNAs targeting three ultra-conserved sequences in the predicted enhancer were designed as follows: CNE-Irx1b-guide1: TCCGTCACGCTGAGATAATC; CNE-Irx1b-guide2: TCAAACACTTTGGGGAACAA; CNE-Irx1b-guide3: TGACCTCTCACCTCGGGCTA. Similarly, three RNAs targeting three ultra-conserved sequences in a random genomic region were designed as follows: Control-guide1: TTGCTTCTGCGCTGAAATAA; Control-guide2: ATGGACTAAAAATTTCACTT; Control-guide3: GAATGTTGATTGTAATTACA. They were purchased from Integrated DNA Technologies as "crRNA", hybridized with their "tracrRNA", forming the guide RNA (gRNA) and incubated with a Cas9 protein (gift from J-P. Concordet). Three nanoliters containing a mix of the 3 resulting ribonucleoproteins (Cas9/gRNA) targeting either the control or the predicted *Irx1b* enhancer were injected at 15µM each.

Acknowledgements

We wish to thank Pierre Vincens for the coordination of computing resources. This work was supported by grants from the French Government and implemented by ANR (ANR-10-BINF-01-03 Ancestrome, ANR–10–LABX–54 MEMOLIFE and ANR–10–IDEX–0001–02 PSL* Research University).

References

1. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15,** 272–286 (2014).

2. Levine, M. Transcriptional enhancers in animal development and evolution. *Curr. Biol.* **20,** R754–63 (2010).
3. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167,** 1170–1187 (2016).
4. Kleinjan, D. A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76,** 8–32 (2005).
5. Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum. Mol. Genet.* **21,** 3255–3263 (2012).
6. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161,** 1012–1025 (2015).
7. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538,** 265–269 (2016).
8. Bal, E. *et al.* Mutations in ACTRT1 and its enhancer RNA elements lead to aberrant activation of Hedgehog signaling in inherited and sporadic basal cell carcinomas. *Nat. Med.* **23,** 1226–1233 (2017).
9. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (2012).
10. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.* **93,** 779–797 (2013).
11. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24,** R102–10 (2015).
12. Ellingsen, S. *et al.* Large-scale enhancer detection in the zebrafish genome. *Development* **132,** 3799–3811 (2005).
13. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3,** e7 (2005).
14. Visel, A. *et al.* Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.* **40,** 158–160 (2008).
15. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).
16. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457,** 854–858 (2009).
17. Blow, M. J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* **42,** 806–810 (2010).
18. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–283 (2011).
19. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
20. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
21. Shibata, Y. *et al.* Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. *PLoS Genet.* **8,** e1002789 (2012).
22. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10,** 1213–1218 (2013).
23. Amano, T. *et al.* Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev. Cell* **16,** 47–57 (2009).
24. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489,** 109–113 (2012).

25. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
26. Jäger, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun* **6,** 6178 (2015).
27. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47,** 598–606 (2015).
28. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25,** 582–597 (2015).
29. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167,** 1369–1384.e19 (2016).
30. Freire-Pritchett, P. *et al.* Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *Elife* **6,** 699 (2017).
31. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).
32. Cao, Q. *et al.* Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genet.* **15,** 272 (2017).
33. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550,** 204–213 (2017).
34. Naville, M. *et al.* Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome. *Nat Commun* **6,** 6904 (2015).
35. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).
36. Cournac, A., Koszul, R. & Mozziconacci, J. The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic Acids Res.* **44,** 245–255 (2016).
37. Bogdanović, O. *et al.* Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. *Genome Res.* **22,** 2043–2053 (2012).
38. Rivera, C. M. & Ren, B. Mapping human epigenomes. *Cell* **155,** 39–55 (2013).
39. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35,** D88–92 (2007).
40. Lee, H. J. *et al.* Developmental enhancers revealed by extensive DNA methylome maps of zebrafish early embryos. *Nat Commun* **6,** 6315 (2015).
41. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502,** 499–506 (2013).
42. Maeso, I., Acemel, R. D. & Gómez-Skarmeta, J. L. Cis-regulatory landscapes in development and evolution. *Curr. Opin. Genet. Dev.* **43,** 17–22 (2017).
43. Bosse, A. *et al.* Identification of the vertebrate Iroquois homeobox gene family with overlapping expression during early development of the nervous system. *Mech. Dev.* **69,** 169–181 (1997).
44. Plessy, C., Dickmeis, T., Chalmel, F. & Strähle, U. Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. *Trends Genet.* **21,** 207–210 (2005).
45. McEwen, G. K. *et al.* Early evolution of conserved regulatory sequences associated with development in vertebrates. *PLoS Genet.* **5,** e1000762 (2009).
46. Visel, A. *et al.* A high-resolution enhancer atlas of the developing telencephalon. *Cell* **152,** 895–908 (2013).
47. Uchikawa, M., Ishida, Y., Takemoto, T., Kamachi, Y. & Kondoh, H. Functional analysis of chicken Sox2 enhancers highlights an array of diverse regulatory elements that are conserved in mammals. *Dev. Cell* **4,** 509–519 (2003).

48. Symon, A. & Harley, V. SOX9: A genomic view of tissue specific expression and action. *Int. J. Biochem. Cell Biol.* **87,** 18–22 (2017).
49. Tschopp, P. & Duboule, D. A genetic approach to the transcriptional regulation of Hox gene clusters. *Annu. Rev. Genet.* **45,** 145–166 (2011).
50. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 1412–1417 (2006).
51. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **5,** e1000598 (2009).
52. Zhu, J., He, F., Hu, S. & Yu, J. On the nature of human housekeeping genes. *Trends Genet.* **24,** 481–484 (2008).
53. Roider, H. G., Lenhard, B., Kanhere, A., Haas, S. A. & Vingron, M. CpG-depleted promoters harbor tissue-specific transcription factor binding signals--implications for motif overrepresentation analyses. *Nucleic Acids Res.* **37,** 6305–6315 (2009).
54. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
55. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).
56. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* **16,** 144–154 (2015).
57. Woolfe, A. & Elgar, G. Organization of conserved elements near key developmental regulators in vertebrate genomes. *Adv. Genet.* **61,** 307–338 (2008).
58. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328,** 1036–1040 (2010).
59. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160,** 554–566 (2015).
60. Liao, B.-Y. & Weng, M.-P. Natural selection drives rapid evolution of mouse embryonic heart enhancers. *BMC Syst Biol* **6 Suppl 2,** S1 (2012).
61. Krijger, P. H. L. & de Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.* **17,** 771–782 (2016).
62. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat. Rev. Genet.* **17,** 661–678 (2016).
63. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485,** 381–385 (2012).
64. Symmons, O. *et al.* Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* **24,** 390–400 (2014).
65. Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149,** 1233–1244 (2012).
66. Symmons, O. *et al.* The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. *Dev. Cell* **39,** 529–543 (2016).
67. Harmston, N. *et al.* Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nat Commun* **8,** 441 (2017).
68. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13,** 103–107 (2003).
69. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14,** 708–715 (2004).
70. Gehrke, A. R. *et al.* Deep conservation of wrist and digit enhancers in fish. *Proc. Natl.*

Acad. Sci. U.S.A. **112,** 803–808 (2015).
71. Bastian, F. *et al.* in *Data Integration in the Life Sciences* **5109,** 124–131 (Springer, Berlin, Heidelberg, 2008).
72. Braasch, I. *et al.* The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat. Genet.* **48,** 427–437 (2016).
73. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **41,** D377–86 (2013).
74. Thomas, P. D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13,** 2129–2141 (2003).
75. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
76. Bessa, J. *et al.* Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Dev. Dyn.* **238,** 2409–2417 (2009).
77. Kawakami, K. *et al.* A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. *Dev. Cell* **7,** 133–144 (2004).
78. Hauptmann, G. & Gerster, T. Two-color whole-mount in situ hybridization to vertebrate and Drosophila embryos. *Trends Genet.* **10,** 266 (1994).