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

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

École Normale Supérieure, PSL Research University, CNRS, Inserm, Institut de Biologie de l'École 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 study gene regulation and enhancer evolution 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 factors and contact promoters in *cis* to activate or repress the transcription of genes into RNA [1]. This control – or regulation – of gene expression by enhancers ensures the fine tuning of protein abundance in cells, in the course of differentiation 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. Disruption of enhancer function has been shown to lead to abnormal gene expression and thus to disease [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 considerations underscore the importance of identifying enhancers and their target genes to better understand genome function.

Numerous methods have been developed to identify enhancers. Some are based on the analysis of the evolutionary 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 have 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, non-coding single nucleotide variants (SNVs) can be associated to significant differences in nearby 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 computational method to identify the target genes of predicted enhancers using signals of evolutionary conserved linkage (or synteny) between putative enhancers and their target genes instead of an experimental assay [34].

This method works in a cell-type or tissue agnostic manner and relies on the analysis of evolutionary signals rather than on a costly and labour intensive experimental setup. 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 predict enhancers in the human and zebrafish genomes as Conserved Non-coding Elements (or CNEs) and applied the PEGASUS method [34] to identify their most likely target genes as the gene(s) in a 1 Mb radius around a predicted enhancer with the most conserved synteny (Figure 1A). This conservation is quantified by a linkage score. We first analysed the UCSC 100-way multiple genome alignment restricted to Sarcopterygii (all vertebrates except fish but including Coelacanth) to identify human CNEs and applied the PEGASUS method, which resulted in predicting 1,311,643 CNEs linked to 18,339 human genes. We also performed 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 analyse the human genome to a situation equivalent to that of zebrafish, we show that this contrast explains the vastly different numbers of CNEs identified in zebrafish (See Discussion & Supplementary material). Therefore, the number of zebrafish CNEs identified here are likely to be a conservative estimate.

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; Figure 1B). Human CNEs cover around 2.35% of the human genome while zebrafish CNEs cover around 0.5% of the zebrafish genome. 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 (Figure 1B), 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 in 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 CNEs are respectively intronic, flanking and jumping (Figure 1B). Interestingly 34% of jumping human CNEs (37% in zebrafish) are located in an intron of a gene that is not their target 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).

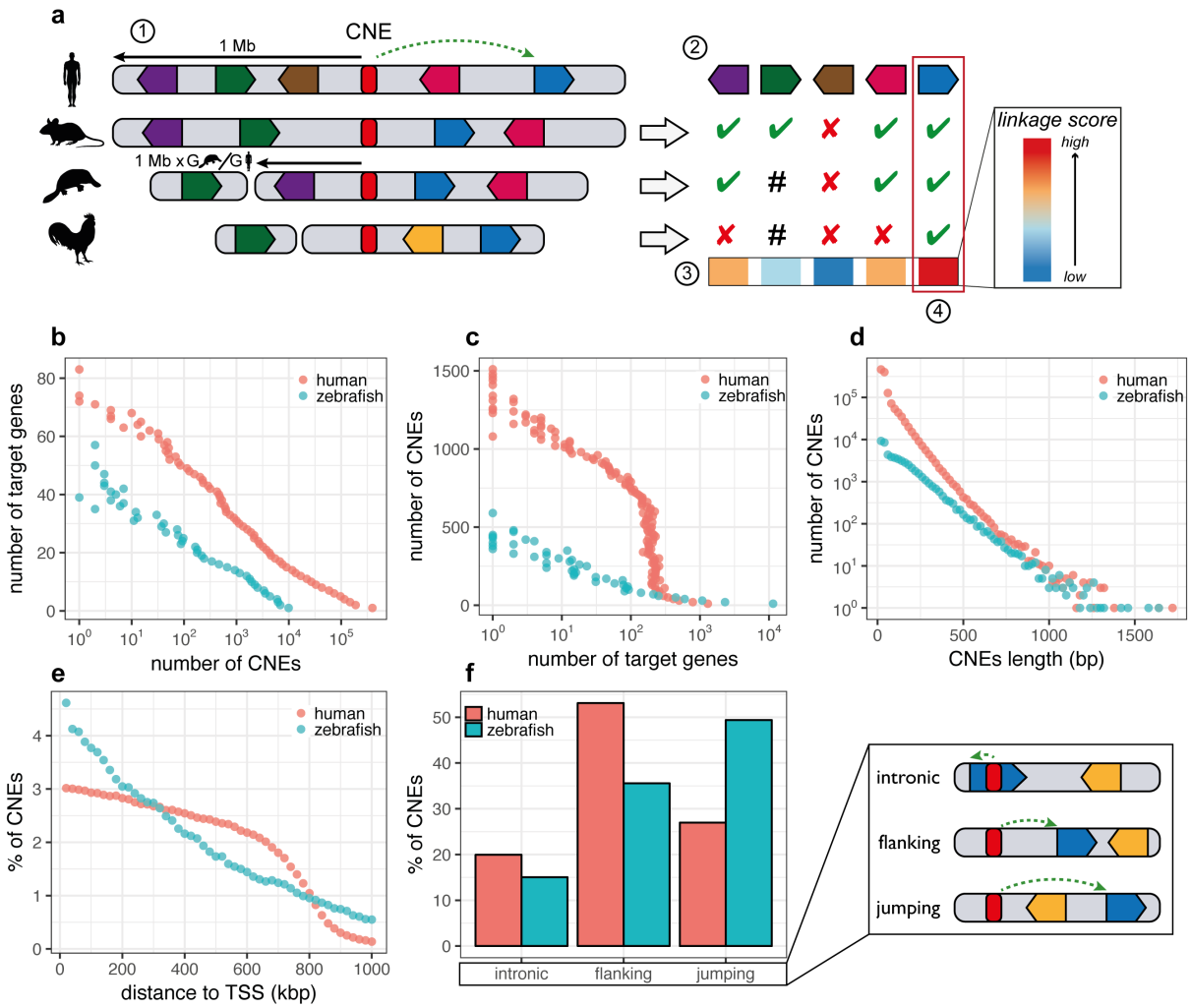

Figure 1: Summary of the PEGASUS method and of results obtained in the human & zebrafish genome.

(a) Schematic summary of the PEGASUS method. 1) After identification of CNES (Conserved Non-coding Elements) in several genomes, all genes in a 1 Mb radius around the CNE in the reference species are considered. 2) For each gene, the method will look in every species where the CNE was defined if the gene is present in the genome & in the radius (scaled by relative genome size, green ticks), present but outside the radius (hash) or absent from the genome (red crosses). Genes are free to move around within this radius. 3) This information is used to compute a linkage score between a CNE & each gene within a 1 Mb radius. 4) The gene(s) with the highest linkage score is(are) considered to be the most probable target(s). (b) Distribution of target genes number per CNE. (c) Distribution of CNE number per target gene. CNE. (d) Distribution of CNE length. (e) Distribution of CNE-target gene TSS distances. (f) proportion of intronic, flanking & jumping CNEs.

From CNEs to predicted enhancers

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 2A & 2B, Supplementary Tables 1 & 2). Interestingly, the two marks associated with active enhancer function (H3K4me1 and K3K27ac [38]) show the strongest overlap (Figure 2). 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 2). In addition, CNEs for which no target gene could be identified have almost no experimental evidence for a regulatory function, as expected if PEGASUS correctly classifies those CNEs has playing no role in regulating the expression of a target gene (Figure 2 and Supplementary Figure 2). 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 2 C & D). Together these results show that human and zebrafish CNEs identified here can be considered as predicted enhancers. We therefore call CNEs predicted enhancers in the remainder of this work.

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. We restricted our analysis to human predicted enhancers linked to a single gene and with a one-to-one overlap with enhancer predictions with other methods (i.e. one and only one CNE overlapping one and only one enhancer region). When confronted to conformation capture based predictions, 28% of our target gene predictions agree with Capture Hi-C data [27,30], but the level of agreement depends on the 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 from the GTEx consortium [33], is higher, with a 57% overlap. Interestingly, for all target gene predictions methods that we analysed, the agreement decreased with distance to the TSS (Supplementary Figure 3), even in the range 0-500 kb where the PEGASUS linkage score is relatively constant (Supplementary Figure 4). This could be explained by the observation that enhancers predicted at longer distances are active in fewer tissues (Supplementary Figure 5), limiting the possibility of overlap with datasets 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 5). We therefore conclude that PEGASUS enhancer-gene predictions are consistent with results from genome wide functional assays. These results are consistent with the fact that PEGASUS, which works in a cell-type or tissue agnostic way, is as powerful in predicting target genes of putative enhancers as labour-intensive methods such as Capture Hi-C. Moreover, using CNEs as putative enhancers enables us to detect more regulatory regions and interactions in a single analysis than experiment-based methods.

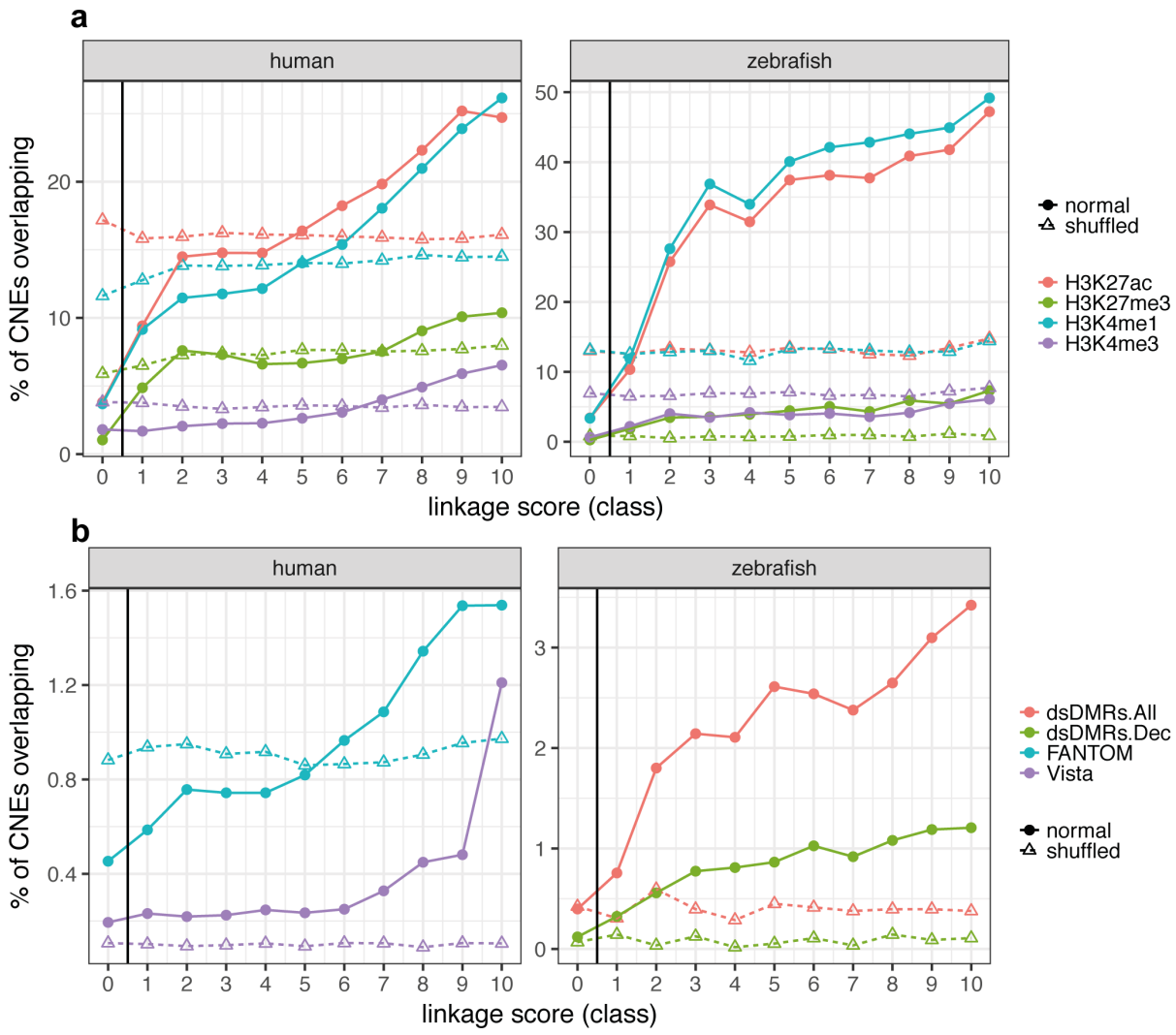

Figure 2: CNEs overlap functional annotations and predicted enhancers.
Percentage of predicted enhancers overlapping chromatin modifications (a) or enhancer predictions (b) in human (left) & zebrafish (right) in classes of linkage score. Predicted enhancers were divided into 10 deciles of equal size, according to their linkage scores. Class 0 represents enhancers with no predicted target gene. The top row 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 row represents overlap with enhancer predictions from FANTOM [31] or Vista [39] in human, & 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 [35], compared to an average of 32% and 41% respectively when we shuffle TAD intervals (proportion test p-values < 10^{-15} for both cell types). Predicted enhancers 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, Wilcoxon rank sum test p-values < 10^{-15} for both cell types, Figure 3), overlapped more with functional marks (14% vs 10% for H3K4me1 & 15% vs 11% for H3K27ac in hESCs, 13% vs 10% for H3K4me1 & 10% vs 15% for H3K27ac in IMR90, all proportion tests p-values < 10^{-15}) and were also closer to each other (median

distance to TSS, 332 kb inside vs 522 kb outside for hESCs, 355 kbp vs 524 kbp for IMR90, Wilcoxon rank sum test p-values < 10^{-15} for both cell types, Figure 3). Finally, we see a striking link between predicted enhancer conservation and their localisation within a TAD. First, human-zebrafish orthologous predicted enhancer-target gene pairs (human-zebrafish orthologous genes with conserved predicted enhancers, see Materials & Methods) are more often located within a TAD than expected by chance (proportion tests p-values < 10^{-15} for both hESCs and IMR90 cells). 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 6). The association between TAD co-localisation and both distance to TSS and conservation depth cannot be explained by chance alone (Supplementary Figure 6). These results, in line with previous observations [42], show evolutionary conservation of linkage between predicted enhancers and their target genes is consistent with topological organisation of chromatin.

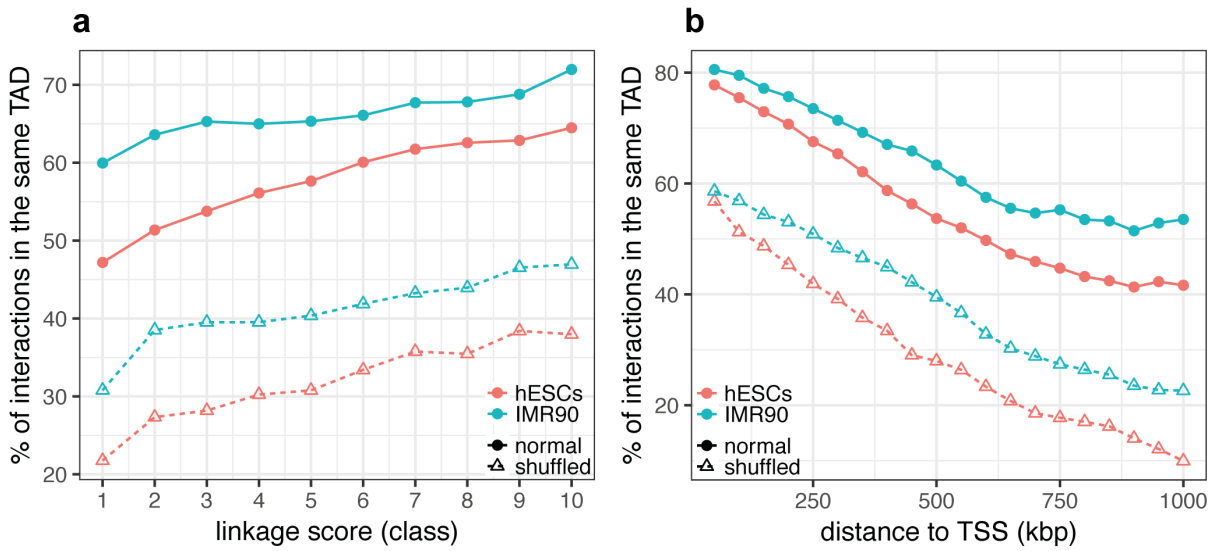

Figure 3: Regulatory interactions are located inside TADs more often than expected by chance.

Percentage of predicted enhancer – target gene interactions co-localisation in the same TAD as a function of (a) linkage score or (b) distance to the TSS. TADs were obtained from two cell types (IMR90 & hESCs [35]). We computed co- localisation percentages in 10 deciles of predicted enhancers based on linkage score, in TADs (labelled "normal") or in TADs after shuffling their localisation (labelled "shuffled").

Function of regulatory interactions conserved in vertebrates

The human and zebrafish predicted enhancer-gene maps described here were built independently using distinct sets of genomes, allowing us to use their intersection as increased evidence for functional and highly conserved enhancer - target gene interactions. This intersection probably outlines those interactions that both species inherited from their last common Euteleostomi ancestor. We identified 1,986 human predicted enhancers and 1,949 zebrafish predicted enhancers 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 1). To illustrate this, we used the CRISPR-Cas9 technology to delete a 236 bp zebrafish predicted enhancer predicted to target the irx1b gene. Its orthologous predicted enhancer is also independently predicted to target IRX1 in the human genome. This gene plays multiple roles during pattern formation of vertebrate embryos [43,44], and we expect

its expression pattern to be tightly regulated by a complex array of enhancers. The deletion of the predicted enhancer greatly decreases the expression of the endogenous gene in several structures of the zebrafish embryo (Figure 4) establishing it as a *bona fide* developmental enhancer. Interestingly, the predicted enhancer targeted by the deletion is closer to another gene (*irx4b*, Figure 4C), yet the expression of this gene is unaffected by the absence of the predicted enhancer (Figure 4D). This further demonstrates that choosing the nearest gene as a putative enhancer's target can lead to false predictions and that PEGASUS can distinguish the correct gene target among closely spaced genes. 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 [12-14,45-47].

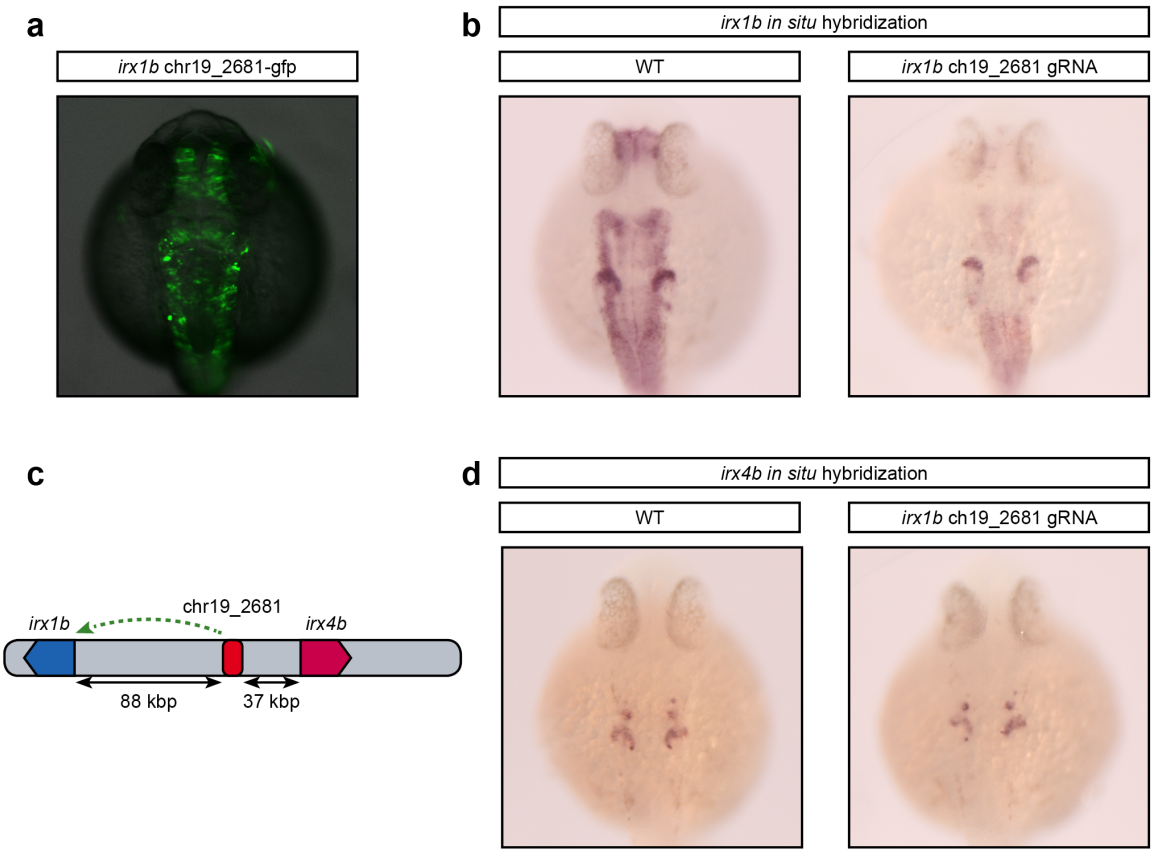

Figure 4: In vivo inactivation of a predicted enhancer targeting irx1b affects its expression.

(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 wild type embryos (WT) or embryos injected with a mix of 3 CRISPR/Cas9 ribonucleoprotein complexes targeted at the predicted *irx1b* enhancer. The CNE activity profile overlaps with *irx1b*'s expression profile, which is comprised of the acousticovestibular ganglia, the caudal diencephalon, the tectum, the hindbrain, the spinal cord & the anterior part of the otic vesicle but not in the mid-hindbrain boundary. *Irx1b*'s expression level is greatly decreased in all these structures when the CRISPR/Cas9 complex is targeted to the predicted enhancer compared to the control, establishing it as a bona fide *irx1b* enhancer. (c) The chr19_2681 CNE, predicted to target *irx1b* is located closer to *irx4b* (37 kbp) than to *irx1b* (88 kbp). (d) In contrast to (b), *irx4b*'s expression profile which includes the anterior part of the otic vesicle & a few cells in the hindbrain is not affected by the CRISPR/Cas9 complex showing that this CNE is specific to *irx1b* & does not regulate *irx4b*.

GO term	fold enrichment	p-value
ventral spinal cord interneuron differentiation	14.02	4.91E-02
positive regulation of cardic muscle cell proliferation	11.38	3.18E-02
central nervous system projection neuron axonogenesis	11.38	3.18E-02
proximal/distal pattern formation	9.35	2.73E-02
cell fate determination	8.90	2.54E-03
forebrain neuron differentiation	6.80	2.77E-02
neuron fate commitment	6.50	4.67E-03
neural crest cell differentiation	6.14	2.37E-02
renal tubule morphogenesis	6.14	2.37E-02
nephron tubule development	5.90	1.26E-02

Table 1: Top 10 overrepresented GO terms in human genes with conserved regulation with zebrafish

Predicted enhancers – 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 [23,24]. Moreover, functional studies have shown that conservation of enhancer function can be independent of sequence conservation [48], that enhancers have high rates of turnover between species [49,50] and that enhancers can change localisation within a TAD without affecting gene regulation [51]. With these results in mind, it is reasonable to speculate that little selective constraint exists 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 predicted enhancers - 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 predicted enhancers and their orthologous target genes scale with intron size (median enhancer-TSS distance ratio = 2.23, median intron length ratio = 2.39, Figure 5). There is therefore no evidence for a constraint on predicted enhancers – gene distance, suggesting that 3D folding principles governing enhancer – promoter contacts are not sensitive to this distance, at least within the limits of the 1 Mb radius adopted here.

Regulation complexity is positively correlated to expression breadth

Enhancers are largely responsible for the spatio-temporal specificity of their target gene's expression (e.g. SOX2 [52], SOX9 [53] or HOX genes [54]). 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. 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 predicted enhancers in the human genome to investigate this question, using expression data from the Bgee database [55]. We observe a positive correlation between the number of enhancers targeting a gene and the number of adult tissues where this gene is expressed (Spearman's $\rho = 0.23$, p-value $< 10^{-15}$, Figure 6). 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 [56-59]. 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 7). Second, we observe that the correlation between predicted enhancer number and expression breadth is stronger for non-CpG genes than for CpG genes (Spearman's ρ = 0.30 & 0.15, respectively, both p-values < 10^{-15}). This lower correlation can be explained by the narrower range of tissue number for CpG genes (Supplementary Figure 8). Interestingly, CpG genes appear to have broad expression regardless of their regulation complexity (Figure 6). Our results show that expression patterns are likely to be driven by regulation complexity, and that the effect is stronger for non CpG genes. Studying the number of adult stages in which a gene is expressed gave almost identical results (Supplementary Figure 9). 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 [55] (Spearman's ρ from -0.08 to 0.15, all p-values <= 0.05). In our study, predicted enhancer number does not seem to predict gene expression levels in the human genome, which is contradictory to numerous studies in mammals & flies [60-63]. However, this may be due to the fact that the transcription levels used for the correlations are measured in specific tissues or cell types, but target gene predictions by PEGASUS are tissue agnostic and not necessarily active in the tissue of interest.

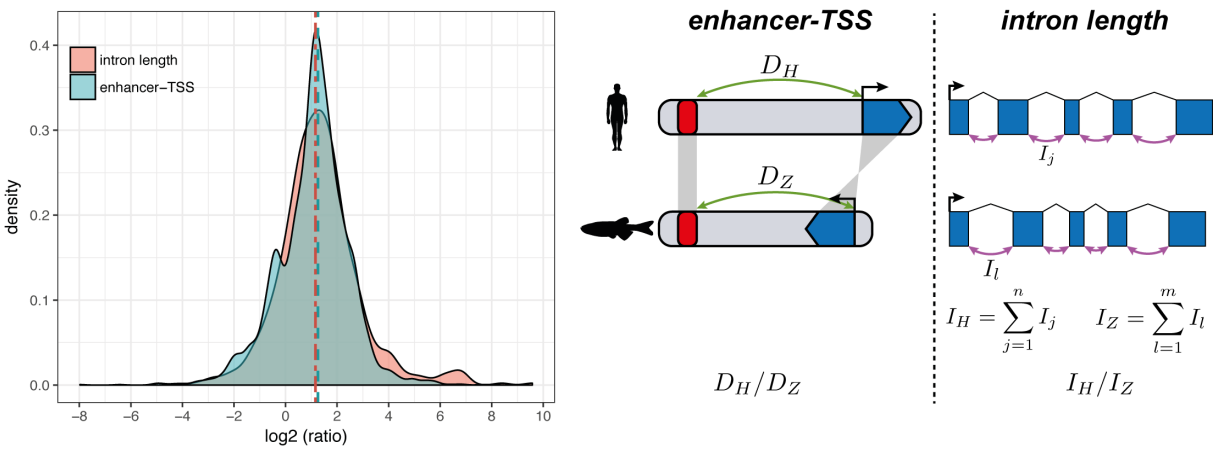

Figure 5: Distances between predicted enhancers and target genes scale with genome size.
Pairwise comparison of enhancer-TSS distances & intron lengths between human & zebrafish. All comparisons were made using the set of human-zebrafish orthologous genes with conserved predicted enhancers. For enhancer-TSS distances, we compared the enhancer-TSS distances (D_H & D_Z for human & zebrafish) for each conserved pair of gene & predicted enhancer. For intron lengths, we compared the total intron length (sum of a gene's intron lengths, I_H & I_Z for human & zebrafish) for each orthologous gene pair. Comparisons were computed as the log2(human / zebrafish) ratios.

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 the

conservation of enhancer-gene linkage follows the 3D organisation of the genome, and that regulatory interaction ancestral to vertebrates concentrate on core functions necessary to build an organism (skeleton, muscles, neurons, vascular system).

Effects of phylogenetic sampling

Enhancer predictions gave contrasted number of enhancers between human and zebrafish (~1,300,000 and ~55,000, respectively). Differences in phylogenetic sampling, i.e. the number of species and their phylogenetic relationships used for enhancer identification and target gene predictions, between the two species may cause this discrepancy. Zebrafish was compared to only 6 other genomes, with zebrafish being an outgroup to all but the spotted gar (Supplementary Figure 1). In contrast, human was compared to 35 other genomes (Supplementary Figure 1). We tested the influence of this phylogenetic sampling by comparing human to 6 species that mirror the phylogenetic relationships in zebrafish (Supplementary Materials). In this reduced set, we could identify approximately 193,000 putative enhancers, a sharp reduction compared to the set benefitting from a full phylogenetic sampling. A limit to our method to identify enhancers and predict target genes is thus the availability of well-annotated and well-reconstructed genomes.

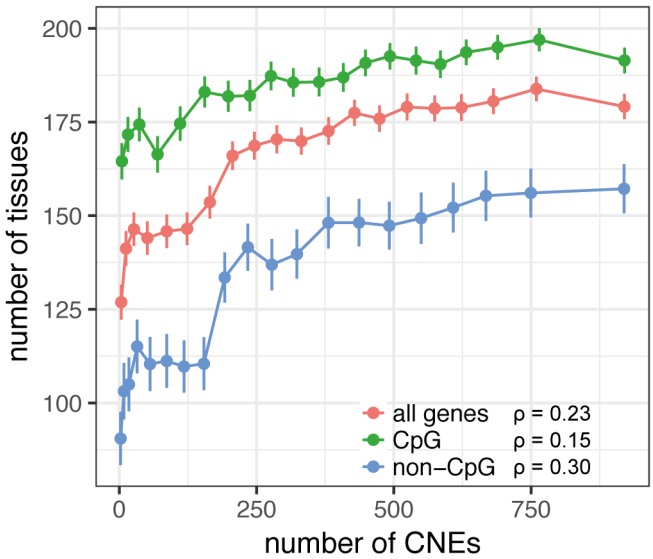

Figure 6: Regulation complexity is positively linked with expression breadth.
Link between the number of predicted enhancers targeting a gene & the number of tissues it is expressed in, for all genes or separating between CpG TSS genes & others. Genes were divided in twenty classes of identical size based on their number of predicted enhancers. Classes were made independently for all genes, for CpG genes & for non-CpG genes. Points & 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.

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 [29,64-66]. 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 [16,67], 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. 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 [49,50,68], it is likely that a fraction of cell-type specific enhancers have had little time to leave detectable footprints of selection in a genome. For the same reasons, PEGASUS will fail to capture species-specific or recently evolved regulatory interactions. 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 (Supplementary Figure 5): short distance enhancers have a regulatory action in more tissues and more likely to have the same predicted target gene when comparing tissues or cell types. 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. TADs are largely conserved between cell types and species [35,36,69]. Enhancer – target genes interactions occur mostly within TADs [70], via DNA looping [71]. 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 [51]. 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 [72], 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 linkage score measuring this conservation.

We identified CNEs in the human and zebrafish genomes in multiple alignments as follows. We first identified seeds of 10bp with at least 9 alignment columns identical between all species considered. These seeds were then extended on both sides, allowing up to three non-conserved alignments columns. We allowed up to 40% of mismatches in a column to consider it as conserved. For the human genome (*hg19* version), we used the UCSC 100-way multiple alignments restricted to 35 Sarcopterygii species (a full list is available in Supplementary Table 3). Alignment blocks had to contain at least 6 species (including human) with one non-primate species to be considered. For the zebrafish genome (*danRer7* version), we generated multiple alignments that include 6 other Neopterygii species (a full list is available in Supplementary Table 4). Multiple alignments were built first by pairwise alignments between zebrafish and other species using LastZ [73], then by using these to build multiple alignments using Multiz [74]. Alignment parameters can be found in Supplementary Materials. Alignments blocks had to contain at least 3 species (including zebrafish) to be considered.

We used PEGASUS [34] to identify target genes in both genomes, considering CNEs as putative enhancers. This method first identifies all protein coding genes in a 1 Mb radius around CNEs. It then computes a linkage score for each gene, reflecting the evolutionary conservation of synteny between a CNE and a particular gene. For each gene around a CNE present in *N* species, the linkage score is computed as follows (equation 1 from [34]):

$$S_L = \sum_{e=1}^{N} S_{e,1} \times R_e - \frac{S_{e,2} + C_e \times (S_{e,3} + S_{e,0})}{R_e}$$

Where C_e is a corrective factor to take assembly errors from low-coverage sequences into account, R_e the rearrangements rate between human or zebrafish and the species e, $S_{e,0}$, $S_{e,1}$, $S_{e,2}$ and $S_{e,3}$ the respective status of the orthologous gene considered in species e (absent or mis-annotated, present and within the correct radius, present and outside the radius, present and on a different chromosome, respectively). The radius in each species is 1Mb corrected by the genome size of species e normalised by the human or zebrafish genome size. Re is computed as (equation 2 in [34]):

$$R_e = \ln\left(\frac{100 \times G}{P_e}\right)$$

Where *G* is the number of gene pairs in the human or zebrafish genome and P_e the number of these pairs that are direct neighbours in species e. The linkage score is then normalised in a [0,1] interval using a sigmoid transformation (equations 3, 4 & 5 in [34]). For a given CNE, the gene with the highest linkage score is defined as its most likely target gene. If more than one gene have the highest linkage score, they all are defined as most likely targets. Adjacent CNEs targeting identical gene, present in the same species, having identical linkage scores

and distant by less than 100bp were merged together. Predicted enhancers located at 100bp or less from an exon were discarded.

445 ## Overlap with functional marks & enhancer predictions

We annotated CNEs that overlapped with functional marks and previous enhancer predictions. We first computed 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 computed the
450 overlap between our predicted enhancers and ATAC-seq peaks in zebrafish [75] and DNase 1 ChIP-seq peaks in human [20]. We finally computed the overlap between our CNEs and other enhancer predictions: we used enhancer predictions from the FANTOM project [31] or from the Vista database [39] for human, and from differentially methylated regions [40] in zebrafish. For all computations, all overlap of at least 1bp were considered. We compared
455 observed overlaps with random overlaps by shuffling region coordinates (ChIP-seq peaks or predicted enhancers) along the human and zebrafish genomes.

Gene expression data

Gene expression values and calls for the human genome were downloaded from the Bgee database [55]. For each gene, we computed the number of adult tissues for human or
460 developmental stage for zebrafish in which a gene is called as expressed. We filtered out terms that had daughter terms for the same gene. Values for gene expression levels (measured in RPKM) were downloaded from the same database. Expression values were available in 21 tissues.

Comparing target gene predictions with chromosome conformation capture
465 predictions

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
470 cells (NECs) [30]. The latter contains interactions present in either cell types or in both simultaneously (labelled "Both"). We considered only CNEs and enhancers with a one to one overlap (one and only one CNE overlapping one and only one enhancer), and both targeting only one gene. In this set of CNEs & enhancers, we computed for each cell type the percentage of PEGASUS predictions and other prediction that agreed with each other.

475 ## Defining orthologous enhancers & target genes between human & zebrafish

We downloaded human-zebrafish and zebrafish-human pairwise chain alignments from 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
480 enhancers targeting orthologous genes.

Because of the evolutionary distance between human and zebrafish, some orthologous regions are difficult to align and are thus impossible to detect. To circumvent this problem, we used the spotted gar genome [76] to identify orthologous predicted enhancers. We downloaded human-spotted gar pairwise chain alignments and used our custom-made
485 zebrafish-spotted gar pairwise chain alignments to respectively map human and zebrafish predicted enhancers onto the spotted gar genome. We considered human and zebrafish 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. Orthologous enhancers identified both directly and via the spotted gar were merged together.

Gene expression enrichment analysis

We performed GO terms enrichment analyses using the PantherDB website [77,78], which performs a Fisher's exact test with FDR multiple test corrections. The test set was defined as human-zebrafish orthologous genes with conserved predicted enhancers defined above. The control set was defined in both species as all genes targeted by at least one predicted enhancer.

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 giving 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 utility available at the UCSC genome browser [79]. For predicted 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 computed random overlap between TADs and regulatory interactions by shuffling the localisation of the TAD domains along the human genome and computing the overlap between predicted enhancers and their targets and these shuffled TADs.

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 [80] upstream of the minimal gata2 promoter/gfp reporter.

Zebrafish egg injections for transgenesis:

The Tol2 transposon/transposase method of transgenesis [81] 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 [82], 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. 2014;15: 272–286. doi:10.1038/nrg3682

2. Levine M. Transcriptional enhancers in animal development and evolution. Curr Biol. 2010;20: R754–63. doi:10.1016/j.cub.2010.06.070

3. Long HK, Prescott SL, Wysocka J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell. 2016;167: 1170–1187. doi:10.1016/j.cell.2016.09.018

4. Kleinjan DA, van Heyningen V. Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet. 2005;76: 8–32. doi:10.1086/426833

5. Smemo S, Campos LC, Moskowitz IP, Krieger JE, Pereira AC, Nobrega MA. Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. Hum Mol Genet. 2012;21: 3255–3263. doi:10.1093/hmg/dds165

6. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161: 1012–1025. doi:10.1016/j.cell.2015.04.004

7. Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature. 2016;538: 265–269. doi:10.1038/nature19800

8. Bal E, Park H-S, Belaid-Choucair Z, Kayserili H, Naville M, Madrange M, 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. 2017;23: 1226–1233. doi:10.1038/nm.4368

9. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337: 1190–1195. doi:10.1126/science.1222794

10. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet. 2013;93: 779–797. doi:10.1016/j.ajhg.2013.10.012

11. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015;24: R102–10. doi:10.1093/hmg/ddv259

12. Ellingsen S, Laplante MA, König M, Kikuta H, Furmanek T, Hoivik EA, et al. Large-scale enhancer detection in the zebrafish genome. Development. 2005;132: 3799–3811. doi:10.1242/dev.01951

13. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, et al. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 2005;3: e7. doi:10.1371/journal.pbio.0030007

14. Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet. 2008;40: 158–160. doi:10.1038/ng.2007.55

15. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011;478: 476–482. doi:10.1038/nature10530

16. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009;457: 854–858. doi:10.1038/nature07730

17. Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-Seq identification of weakly conserved heart enhancers. Nat Genet. 2010;42: 806–810. doi:10.1038/ng.650

18. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. A unique chromatin signature uncovers early developmental enhancers in humans. Nature. 2011;470: 279–283. doi:10.1038/nature09692

19. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473: 43–49. doi:10.1038/nature09906

20. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74. doi:10.1038/nature11247

21. Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, Tewari AK, et al. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. PLoS Genet. 2012;8: e1002789. doi:10.1371/journal.pgen.1002789

22. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10: 1213–1218. doi:10.1038/nmeth.2688

23. Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet. 2003;12: 1725–1735.

24. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489: 109–113. doi:10.1038/nature11279

25. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326: 289–293. doi:10.1126/science.1181369

26. Jäger R, Migliorini G, Henrion M, Kandaswamy R, Speedy HE, Heindl A, et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. Nat Commun. 2015;6: 6178. doi:10.1038/ncomms7178

27. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet. 2015;47: 598–606. doi:10.1038/ng.3286

28. Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre B-M, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. 2015;25: 582–597. doi:10.1101/gr.185272.114

29. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016;167: 1369–1384.e19. doi:10.1016/j.cell.2016.09.037

30. Freire-Pritchett P, Schoenfelder S, Várnai C, Wingett SW, Cairns J, Collier AJ, et al. Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. Elife. 2017;6: 699. doi:10.7554/eLife.21926

31. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507: 455–461. doi:10.1038/nature12787

32. Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, et al. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. Nat Genet. 2017;15: 272. doi:10.1038/ng.3950

33. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550: 204–213. doi:10.1038/nature24277

34. Naville M, Ishibashi M, Ferg M, Bengani H, Rinkwitz S, Krecsmarik M, et al. Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome. Nat Commun. 2015;6: 6904. doi:10.1038/ncomms7904

35. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485: 376–380. doi:10.1038/nature11082

36. Cournac A, Koszul R, Mozziconacci J. The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. Nucleic Acids Res. 2016;44: 245–255. doi:10.1093/nar/gkv1292

37. Bogdanović O, Fernández-Miñán A, Tena JJ, la Calle-Mustienes de E, Hidalgo C, van Kruysbergen I, et al. Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. Genome Res. 2012;22: 2043–2053. doi:10.1101/gr.134833.111

38. Rivera CM, Ren B. Mapping human epigenomes. Cell. 2013;155: 39–55. doi:10.1016/j.cell.2013.09.011

39. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res. 2007;35: D88–92. doi:10.1093/nar/gkl822

40. Lee HJ, Lowdon RF, Maricque B, Zhang B, Stevens M, Li D, et al. Developmental enhancers revealed by extensive DNA methylome maps of zebrafish early embryos. Nat Commun. 2015;6: 6315. doi:10.1038/ncomms7315

41. de Laat W, Duboule D. Topology of mammalian developmental enhancers and their regulatory landscapes. Nature. 2013;502: 499–506. doi:10.1038/nature12753

42. Maeso I, Acemel RD, Gómez-Skarmeta JL. Cis-regulatory landscapes in development and evolution. Curr Opin Genet Dev. 2017;43: 17–22. doi:10.1016/j.gde.2016.10.004

43. Bosse A, Zülch A, Becker MB, Torres M, Gómez-Skarmeta JL, Modolell J, et al. Identification of the vertebrate Iroquois homeobox gene family with overlapping expression during early development of the nervous system. Mech Dev. 1997;69: 169–181.

44. Lecaudey V, Anselme I, Dildrop R, Rüther U, Schneider-Maunoury S. Expression of the zebrafish Iroquois genes during early nervous system formation and patterning. J Comp Neurol. Wiley Subscription Services, Inc., A Wiley Company; 2005;492: 289–302. doi:10.1002/cne.20765

45. Plessy C, Dickmeis T, Chalmel F, Strähle U. Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. Trends Genet. 2005;21: 207–210. doi:10.1016/j.tig.2005.02.006

46. McEwen GK, Goode DK, Parker HJ, Woolfe A, Callaway H, Elgar G. Early evolution of conserved regulatory sequences associated with development in vertebrates. PLoS Genet. 2009;5: e1000762. doi:10.1371/journal.pgen.1000762

47. Visel A, Taher L, Girgis H, May D, Golonzhka O, Hoch RV, et al. A high-resolution enhancer atlas of the developing telencephalon. Cell. 2013;152: 895–908. doi:10.1016/j.cell.2012.12.041

48. Domené S, Bumaschny VF, de Souza FSJ, Franchini LF, Nasif S, Low MJ, et al. Enhancer turnover and conserved regulatory function in vertebrate evolution. Philos Trans R Soc Lond, B, Biol Sci. 2013;368: 20130027–20130027. doi:10.1098/rstb.2013.0027

49. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science. 2010;328: 1036–1040. doi:10.1126/science.1186176

50. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, et al. Enhancer evolution across 20 mammalian species. Cell. 2015;160: 554–566. doi:10.1016/j.cell.2015.01.006

51. Symmons O, Pan L, Remeseiro S, Aktas T, Klein F, Huber W, et al. The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. Dev Cell. 2016;39: 529–543. doi:10.1016/j.devcel.2016.10.015

52. 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. 2003;4: 509–519.

53. Symon A, Harley V. SOX9: A genomic view of tissue specific expression and action. Int J Biochem Cell Biol. 2017;87: 18–22. doi:10.1016/j.biocel.2017.03.005

54. Tschopp P, Duboule D. A genetic approach to the transcriptional regulation of Hox gene clusters. Annu Rev Genet. 2011;45: 145–166. doi:10.1146/annurev-genet-102209-163429

55. Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, Robinson-Rechavi M. Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. Data Integration in the Life Sciences. Berlin, Heidelberg: Springer, Berlin, Heidelberg; 2008. pp. 124–131. doi:10.1007/978-3-540-69828-9_12

56. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci USA. 2006;103: 1412–1417. doi:10.1073/pnas.0510310103

57. Ramsköld D, Wang ET, Burge CB, Sandberg R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol. 2009;5: e1000598. doi:10.1371/journal.pcbi.1000598

58. Zhu J, He F, Hu S, Yu J. On the nature of human housekeeping genes. Trends Genet. 2008;24: 481–484. doi:10.1016/j.tig.2008.08.004

59. Roider HG, Lenhard B, Kanhere A, Haas SA, Vingron M. CpG-depleted promoters harbor tissue-specific transcription factor binding signals--implications for motif overrepresentation analyses. Nucleic Acids Res. 2009;37: 6305–6315. doi:10.1093/nar/gkp682

60. Odom DT, Dowell RD, Jacobsen ES, Nekludova L, Rolfe PA, Danford TW, et al. Core transcriptional regulatory circuitry in human hepatocytes. Mol Syst Biol. 2006;2: 2006.0017. doi:10.1038/msb4100059

61. Mikkelsen TS, Xu Z, Zhang X, Wang L, Gimble JM, Lander ES, et al. Comparative epigenomic analysis of murine and human adipogenesis. Cell. 2010;143: 156–169. doi:10.1016/j.cell.2010.09.006

62. Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013;339: 1074–1077. doi:10.1126/science.1232542

63. Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. Nat Ecol Evol. 2018;2: 152–163. doi:10.1038/s41559-017-0377-2

64. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518: 317–330. doi:10.1038/nature14248

65. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159: 1665–1680. doi:10.1016/j.cell.2014.11.021

66. Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol. 2015;16: 144–154. doi:10.1038/nrm3949

67. Woolfe A, Elgar G. Organization of conserved elements near key developmental regulators in vertebrate genomes. Adv Genet. Elsevier; 2008;61: 307–338. doi:10.1016/S0065-2660(07)00012-0

68. Liao B-Y, Weng M-P. Natural selection drives rapid evolution of mouse embryonic heart enhancers. BMC Syst Biol. 2012;6 Suppl 2: S1. doi:10.1186/1752-0509-6-S2-S1

69. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;485: 381–385. doi:10.1038/nature11049

70. Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, et al. Functional and topological characteristics of mammalian regulatory domains. Genome Res. 2014;24: 390–400. doi:10.1101/gr.163519.113

71. Deng W, Lee J, Wang H, Miller J, Reik A, Gregory PD, et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell. 2012;149: 1233–1244. doi:10.1016/j.cell.2012.03.051

72. Harmston N, Ing-Simmons E, Tan G, Perry M, Merkenschlager M, Lenhard B. Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. Nat Commun. 2017;8: 441. doi:10.1038/s41467-017-00524-5

73. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, et al. Human-mouse alignments with BLASTZ. Genome Res. 2003;13: 103–107. doi:10.1101/gr.809403

74. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 2004;14: 708–715. doi:10.1101/gr.1933104

75. Gehrke AR, Schneider I, la Calle-Mustienes de E, Tena JJ, Gomez-Marin C, Chandran M, et al. Deep conservation of wrist and digit enhancers in fish. Proc Natl Acad Sci USA. 2015;112: 803–808. doi:10.1073/pnas.1420208112

76. Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, et al. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat Genet. 2016;48: 427–437. doi:10.1038/ng.3526

77. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 2013;41: D377–86. doi:10.1093/nar/gks1118

78. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13: 2129–2141. doi:10.1101/gr.772403

79. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12: 996–1006. doi:10.1101/gr.229102

80. Bessa J, Tena JJ, la Calle-Mustienes de E, Fernández-Miñán A, Naranjo S, Fernández A, 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. 2009;238: 2409–2417. doi:10.1002/dvdy.22051

81. Kawakami K, Takeda H, Kawakami N, Kobayashi M, Matsuda N, Mishina M. A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. Dev Cell. 2004;7: 133–144. doi:10.1016/j.devcel.2004.06.005

82. Hauptmann G, Gerster T. Two-color whole-mount in situ hybridization to vertebrate and Drosophila embryos. Trends Genet. 1994;10: 266.