



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

## Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome

### Citation for published version:

Naville, M, Ishibashi, M, Ferg, M, Bengani, H, Rinkwitz, S, Krecsmarik, M, Hawkins, TA, Wilson, SW, Manning, E, Chilamakuri, CSR, Wilson, DI, Louis, A, Raymond, FL, Rastegar, S, Straehle, U, Lenhard, B, Bally-Cuif, L, van Heyningen, V, FitzPatrick, DR, Becker, TS & Crollius, HR 2015, 'Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome' Nature Communications, vol. 6, 6904. DOI: 10.1038/ncomms7904

### Digital Object Identifier (DOI):

[10.1038/ncomms7904](https://doi.org/10.1038/ncomms7904)

### Link:

[Link to publication record in Edinburgh Research Explorer](#)

### Document Version:

Publisher's PDF, also known as Version of record

### Published In:

Nature Communications

### Publisher Rights Statement:

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

### General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

### Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



ARTICLE

Received 1 May 2014 | Accepted 12 Mar 2015 | Published 24 Apr 2015

DOI: 10.1038/ncomms7904

OPEN

# Long-range evolutionary constraints reveal *cis*-regulatory interactions on the human X chromosome

Magali Naville<sup>1,2,3,\*</sup>, Minaka Ishibashi<sup>4,\*</sup>, Marco Ferg<sup>5,\*</sup>, Hemant Bengani<sup>6,\*</sup>, Silke Rinkwitz<sup>4</sup>, Monika Krecsmarik<sup>7</sup>, Thomas A. Hawkins<sup>8</sup>, Stephen W. Wilson<sup>8</sup>, Elizabeth Manning<sup>4</sup>, Chandra S.R. Chilamakuri<sup>9</sup>, David I. Wilson<sup>10</sup>, Alexandra Louis<sup>1,2,3</sup>, F. Lucy Raymond<sup>11</sup>, Sepand Rastegar<sup>5</sup>, Uwe Strähle<sup>5</sup>, Boris Lenhard<sup>12</sup>, Laure Bally-Cuif<sup>7</sup>, Veronica van Heyningen<sup>6</sup>, David R. FitzPatrick<sup>6</sup>, Thomas S. Becker<sup>4,13,\*\*</sup>  
& Hugues Roest Crollius<sup>1,2,3,\*\*</sup>

Enhancers can regulate the transcription of genes over long genomic distances. This is thought to lead to selection against genomic rearrangements within such regions that may disrupt this functional linkage. Here we test this concept experimentally using the human X chromosome. We describe a scoring method to identify evolutionary maintenance of linkage between conserved noncoding elements and neighbouring genes. Chromatin marks associated with enhancer function are strongly correlated with this linkage score. We test >1,000 putative enhancers by transgenesis assays in zebrafish to ascertain the identity of the target gene. The majority of active enhancers drive a transgenic expression in a pattern consistent with the known expression of a linked gene. These results show that evolutionary maintenance of linkage is a reliable predictor of an enhancer's function, and provide new information to discover the genetic basis of diseases caused by the mis-regulation of gene expression.

<sup>1</sup>Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, 46 rue d'Ulm, Paris F-75005, France. <sup>2</sup>CNRS, UMR 8197, Paris F-75005, France. <sup>3</sup>Inserm, U1024, Paris F-75005, France. <sup>4</sup>Brain and Mind Research Institute, Sydney Medical School, University of Sydney, Camperdown, New South Wales 2050, Australia. <sup>5</sup>Institute of Toxicology and Genetics and European Zebrafish Resource Centre, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany. <sup>6</sup>MRC Human Genetics Unit, MRC Institute of Medical Genetic and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK. <sup>7</sup>Paris-Saclay Institute for Neuroscience (Neuro-PSI), UMR9197 CNRS-Université Paris Sud, Avenue de la Terrasse, Gif-sur-Yvette 91190, France. <sup>8</sup>C.D.B. Division of Biosciences, Anatomy building, UCL, Gower street, London, WC1E 6BT, UK. <sup>9</sup>Department of Tumor Biology, The Norwegian Radium Hospital, 0310 Oslo, Norway. <sup>10</sup>University of Southampton and University Hospital Southampton NHS Foundation Trust, Centre for Human Development, Stem Cells and Regeneration, MP808, Faculty of Medicine, Southampton General Hospital, Tremona Road, Southampton 16 6YD, UK. <sup>11</sup>Cambridge Institute for Medical Research, University of Cambridge, Hills Road, Cambridge CB2 0XY, UK. <sup>12</sup>Institute of Clinical Sciences, MRC Clinical Sciences Centre, Faculty of Medicine, Imperial College London, Hammersmith Hospital Campus, Du Cane Road, London W12 0NN, UK. <sup>13</sup>Department of Clinical Medicine, University of Bergen, Bergen 5009, Norway. \* These authors contributed equally to this work. \*\* These authors jointly supervised this work. Correspondence and requests for materials should be addressed to T.S.B. (email: tom.becker@sydney.edu.au) or to H.R.C. (email: hrc@ens.fr).

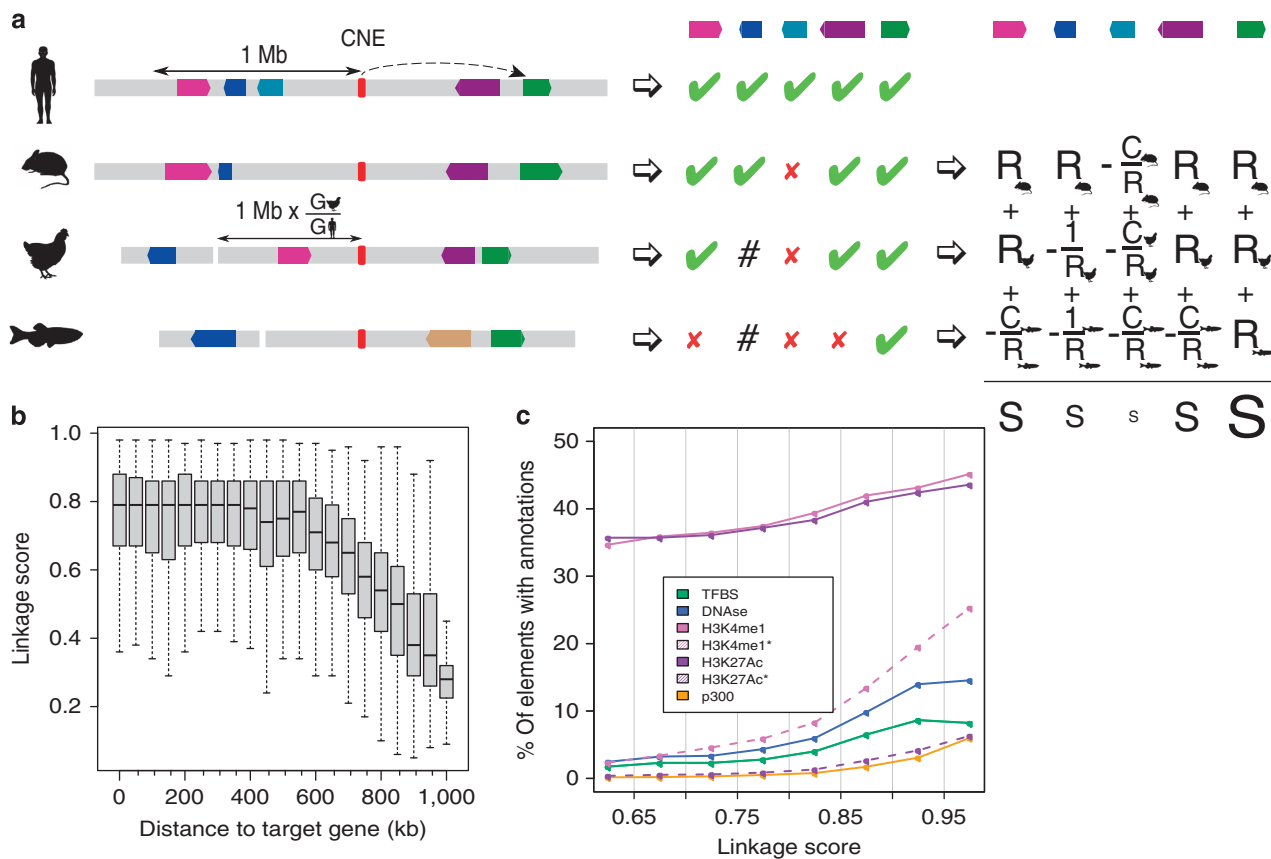
**C**is-regulation is a vital mechanism for the normal development and health of an organism. The cis-regulation of protein-coding gene expression in vertebrate genomes is mediated by regulatory factors binding to enhancer elements that may be located as much as 1.5 Mb from their target genes<sup>1,2</sup>, and longer distances are entirely possible. Given the importance of this cis-interaction, negative selection is thought to prevent the evolutionary fixation of rearrangements that would either physically dissociate the enhancer from the target gene or separate them by an excessive genomic distance. Genomic regions bearing these properties have been described as genome regulatory blocks<sup>3,4</sup>, but systematic efforts to exploit this evolutionary signature on a genomic scale<sup>5</sup> have yet to be experimentally validated. Here we perform such an analysis on the human X chromosome, by developing a score that measures the evolutionary linkage between putative enhancers and their surrounding genes. We show that conserved noncoding elements (CNEs) showing the highest linkage scores are also enriched in functional marks such as epigenetic modifications characteristic of enhancers. We experimentally test >1,000 CNEs for their ability to replicate the expression pattern of their most strongly linked genes, and validate the predicted association for 60% of the cases where the expression pattern of the target gene was known. We finally show that putative enhancers linked to the same target gene are enriched in sequence motifs that may trigger the binding of specific transcription factors.

## Results

**Prediction of CNE/target gene associations.** We identified human X-chromosome CNEs by scanning a multispecies genomic alignment encompassing 46 vertebrate genomes<sup>6</sup> (Supplementary Fig. 1A), and looked for conserved regions, excluding exons and repeat sequences (Methods). This set was then merged with CNEs previously identified in eutherian mammals<sup>7</sup>. Together, these regions represent 174,473 distinct CNEs covering 4.4% of the human X chromosome, likely to represent most noncoding sequences under conservation. To test the hypothesis that functional interactions translate in physical linkage, we first devised a scoring procedure based on evolutionary conservation of linkage between a CNE and one of the human genes located within a radius of 1 Mb from the CNE. For a given CNE, the position of the orthologous CNEs were first sought in all the vertebrate genomes that align at this position. Next, the orthologs of the human genes found in the 1-Mb radius were also collected in all vertebrate genomes. Four situations may arise depending on whether and where the orthologous gene is present: (i) it too is linked to the orthologous CNE in the defined radius, (ii) it is located on the same chromosome but beyond the defined radius, (iii) it is located on a different chromosome and (iv) it is not annotated in the genome. In each genome, each situation was diagnosed and labelled with a score that accounts for the conservation of synteny between the human genome and the genome of interest, and the sequencing coverage of the latter (Fig. 1a and Methods). The maximum genomic interval allowed for linking the orthologous CNE and gene(s) was conservatively taken as 1 Mb but scaled in each genome depending on its relative size compared with the human genome. Together, this linkage and synteny information was used to compute an absolute score  $S_A$  between each CNE and each human gene within the 1 Mb radius ( $0 < S_A < 1$ ), reflecting the degree of linkage between them in vertebrate genomes (Fig. 1a and Methods). For each CNE, the best scoring genes were selected as plausible targets, with no minimal score threshold (Supplementary Data 1), and CNEs targeting the same genes were merged if their positions were

< 100 bp apart (Supplementary Fig. 1B). These merged CNEs are hereafter called RegHsa elements. We identified 102,647 RegHsa elements on the X chromosome with a mean size of 88 bp. Only 1% of RegHsas are not associated with a potential target gene (that is, their distance to the nearest human gene exceeds 1 Mb), 37.5% are associated with a single predicted target (single targets), and 61.5% are associated with several target genes with identical maximal score (multiple targets, not necessarily contiguous). Such multiple targets occur when evolutionarily neutral breakpoints have not yet dissociated the locus, some 'bystander' genes may be captured in a genome regulatory block between an enhancer and its target gene<sup>4</sup>, or an enhancer may regulate several neighbouring genes. Of the 812 protein-coding genes annotated on the X chromosome, 389 were associated with at least one RegHsa element, while some genes, including *DIAP2*, *DMD* or *ODZ1*, are associated with > 100 RegHsa elements. Of the RegHsa elements predicted to target a single gene, 60.7% target a gene that is not their direct neighbour. Interestingly, we observe a remarkably stable median linkage score in a 600-kb radius from the RegHsa element, with a sharp drop in linkage score values beyond this distance (Fig. 1b). Although enhancers are known to function beyond 600 kb, this result may indicate that factors such as the three-dimensional chromatin conformation or breakpoint frequencies may generally be unfavourable to long-range regulatory interactions beyond this distance.

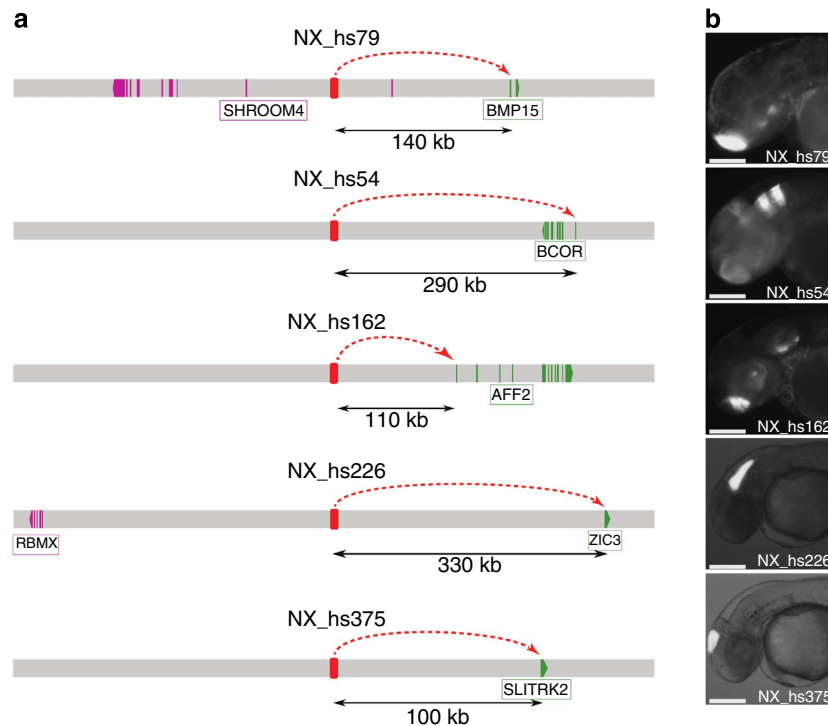
**The linkage score is correlated with functional marks.** If our method correctly reflects a functional association between enhancers and their target genes, we expect the linkage score  $S_A$  to correlate with functional annotations known to be associated with enhancers. To examine this, we annotated all CNEs that constitute RegHsa elements with functional signals known to be associated with enhancer function including chromatin accessibility by DNaseI assays, H3K4me1, H3K4me3, H3K27ac histone modifications and transcription factor-binding assays obtained from seven human cell lines<sup>8</sup>, as well as p300 signals from the mouse embryonic heart, forebrain, midbrain and limb<sup>9,10</sup>. Because the human X chromosome is known to harbour a high proportion of genes involved in cognitive functions and expressed in neural tissues<sup>11</sup>, we also performed H3K4me1, H3K27ac and p300 ChIP-on-chip experiments on human foetal brain and mouse E14.5, E16.5 and P0 developing brain tissues (Methods). When ranking CNEs and target gene associations by increasing the  $S_A$  score, we observe a pronounced enrichment in all functional annotations (Fig. 1c and Methods), with a fivefold increase in DNaseI accessibility (average over seven human cell lines) and a striking 10.8-fold increase in H3K4me1 marks in human developing brain. Notably, the enrichment is not solely a consequence of the positive correlation between linkage score and conservation (Supplementary Fig. 2) because the result remains even when controlling for conservation (Supplementary Fig. 3). High scoring RegHsa elements ( $S_A > 0.9$ ) are linked to genes showing a marked enrichment in gene ontology (GO) terms, notably those associated to neuronal cell body, axon guidance and synapse (Supplementary Table 1). Finally, the linkage score  $S_A$  strongly correlates with an enrichment of known transcription factor-binding motifs (Supplementary Fig. 4). Together, these results indicate that cis-interactions predicted only using evolutionary information are enriched in functional enhancers. Notably, this result is not limited to the X chromosome, because when we compute the  $S_A$  score on autosomes, they also show the same enrichment in functional annotations as a function of linkage score (Supplementary Fig. 5).



**Figure 1 | Scoring evolutionary linkage.** (a) Strategy to compute the linkage score. The presence of human genes in a 1-Mb radius around a CNE are recorded, as well as the simultaneous presence/absence of their orthologs in the vicinity of the orthologous CNEs in different species (green ticks/red crosses, respectively, in the middle panel; hash signs indicate genes located beyond the 1 Mb threshold). The presence of an orthologue is weighted by the degree of conserved synteny  $R$  between this genome and the human genome, while the costs for the absence of a gene account for the sequencing coverage  $C$  of the genome. The final linkage score  $S$  is the sum of these weights in the different genomes where the CNE is present (right panel). The gene(s) showing the maximum linkage score to a given CNE is considered to be the most likely target. (b) The linkage score of the CNE-target predictions were grouped in bins according to the genomic distance between the CNE and its predicted target ( $x$  axis). The median linkage score of the distributions ( $y$  axis) is stable for genes located up to  $\sim 600$  kb from the RegHsa element. (c) The linkage score is strongly correlated with an enrichment in annotations linked to enhancer function. An asterisk indicates data generated during this project.

**Functional validation of predicted interactions.** Next we directly tested the enhancer function of the interaction predicted by our comparative and functional genomic analyses by using transgenic assays. We selected 450 regions of  $\sim 1$  kb on the human X chromosome and overlapping 1,013 human RegHsa elements. These elements encompass a range of conservation levels and a large range of  $S_A$  scores (0.320–0.980) linking them to genes known to be involved in brain development (Supplementary Data 2). We examined their ability to drive specific green fluorescent protein (GFP) expression patterns in zebrafish embryos, by analysing at least five different insertions in F1 lines at 2 days post fertilization for each element. RegHsa elements with a reproducible or partially reproducible pattern of expression (448 cases) allowed us to test if the predicted target gene or genes of the enhancer are compatible with this pattern. For 323 RegHsas, expression data were available for the zebrafish (described in the ZFIN database<sup>12</sup>) for at least one predicted target. Of these, 200 RegHsa elements (60%) drive a transgenic GFP pattern that fully or partially overlaps the ZFIN pattern of one of the predicted targets (Fig. 2, Supplementary Figs 6 and 7, Supplementary Data 2 and Methods). These cases support the prediction that the enhancer indeed regulates the target gene showing the best  $S_A$  score. Consistent with this result, the average  $S_A$  score is significantly higher for the 200 supported

enhancer–gene associations than for those that are not ( $S_A$  score 0.923 versus 0.863;  $P < 2.10^{-16}$ , Wilcoxon test). Interestingly, while 25% of tested RegHsa elements are conserved in zebrafish genomic DNA, this figure increases to 44% for elements with a predicted target that is supported in the transgenic experiments. This shows that conservation of a RegHsa elements is correlated with its functional property as enhancer, but it also shows that absence of conservation in fish does not preclude validation since 56% of enhancers are validated without conservation in fish. To further confirm the identity of the target gene in a limited number of cases, we verified if the enhancer drives GFP expression in the same brain region or cell type where the mRNA of its predicted target gene is expressed. To this end, we performed a detailed anatomical characterization of the GFP expression pattern in juvenile and/or adult zebrafish brains, for transgenic lines corresponding to 15 different human sequences overlapping 67 RegHsas (Methods). Out of the 15 transgenic assays analysed, 13 (87%) show that the gene that is evolutionarily linked to the RegHsa element is expressed in a pattern that completely (6 cases) or partially (7 cases) overlaps with the transgenic GFP pattern in either juvenile or adult zebrafish brain (Supplementary Data 3). For example, the RegHsa0032185 element is predicted to regulate the *BCOR* gene ( $S_A = 0.917$ ) yet is located 286 Kb downstream of the nearest *BCOR* promoter



**Figure 2 | Cis-regulatory interactions predicted by the linkage score are experimentally tested in developing zebrafish.** (a) Individual exons of the predicted target gene are depicted in green and of neighbouring genes in pink. The arrowhead indicates the direction of transcription. Distance in kilobases between the CNE and the promoter of the predicted gene are indicated. (b) The predictions are supported by transgenic analysis in zebrafish. Expression at 48 hpf: NX\_hs79: telencephalon (scale bar, 125  $\mu$ m); NX\_hs54: hindbrain, telencephalon (scale bar, 125  $\mu$ m); NX\_hs162: telencephalon, hypothalamus, otic vesicle (scale bar, 125  $\mu$ m); NX\_hs226: hindbrain (scale bar, 200  $\mu$ m); NX\_hs375: midbrain (scale bar, 200  $\mu$ m).

(Fig. 2a). The elements reproducibly drive GFP expression in the developing zebrafish telencephalon and hindbrain. Neuroanatomical characterization of GFP expression in transgenic zebrafish lines carrying the RegHsa230032185 compared with endogenous zebrafish *bcor* mRNA expression in both juvenile and adult brains shows a strong overlap in the anterior telencephalon (Fig. 3). Critically, the GFP expression pattern strongly overlaps the endogenous zebrafish *bcor* mRNA expression (Fig. 3c). In addition, target gene predictions are consistent with published chromatin interaction maps. Indeed, of the 2,096 RegHsa elements that overlap the regions involved in 781 long-range chromatin interactions experimentally observed on the X chromosome by ChIA-PET in five human cell lines<sup>13</sup>, 69% are evolutionary associated (that is, show the best  $S_A$  score) with the same gene as shown to be involved in the chromatin interaction ( $P$  value  $< 10^{-5}$ , permutation test). Notably, this overlap is the same if we only consider cases where the predicted target is the nearest gene to the RegHsa element or if we consider cases where one or more genes separate the two. Together, these results support the original target gene prediction, which was obtained solely using genome comparisons. Interestingly, while our data agree with the 'nearest gene' strategy 60% of the time (as does the ChIA-PET data, 62%), a greater rate of validation is observed when comparing our data with the ChIA-PET data (69%), which necessarily includes non-nearest genes.

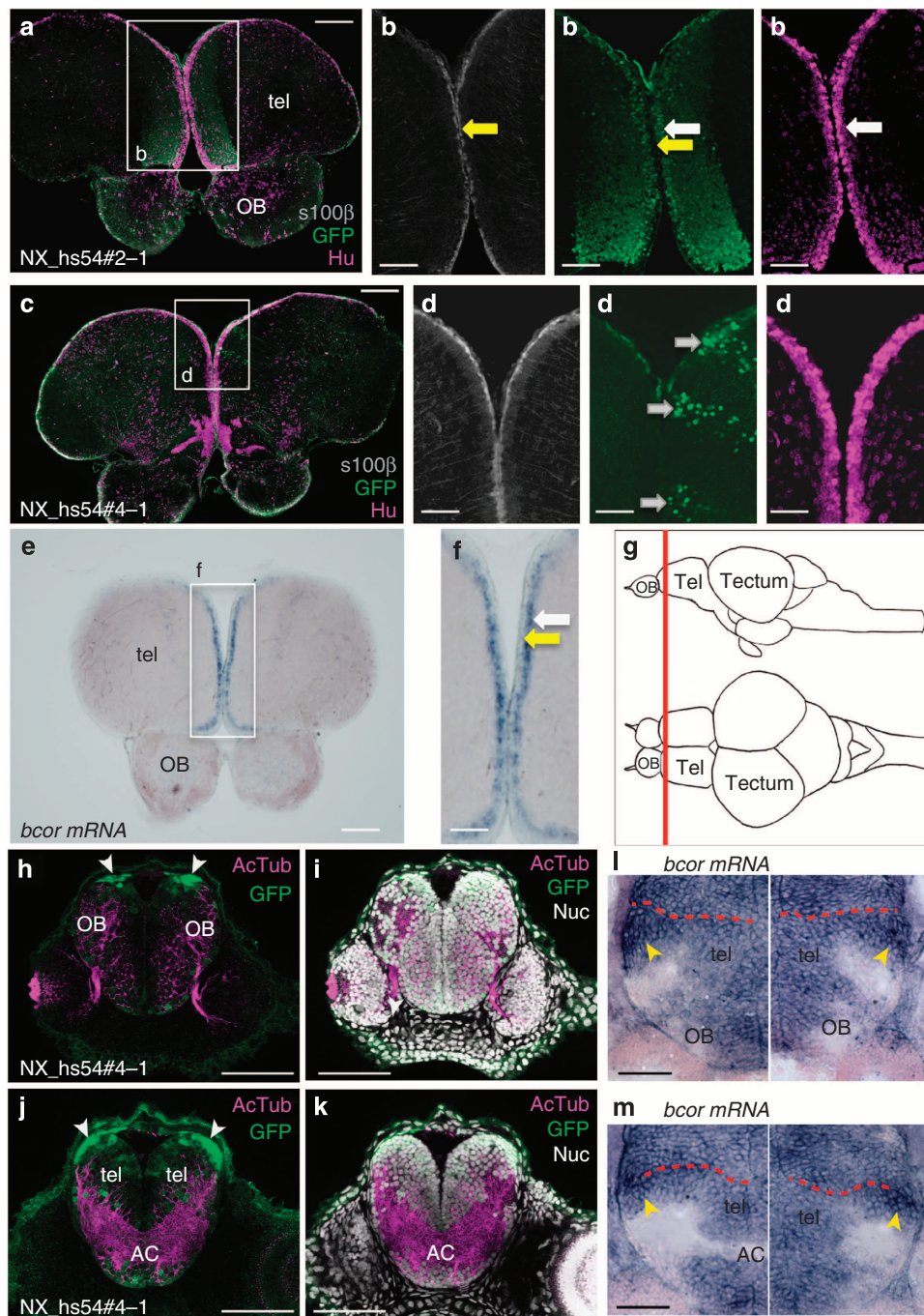
**Motif discovery in CNEs assigned to the same target gene.** On average, 389 single target genes are associated with a mean of 17 RegHsa elements each with  $S_A > 0.9$ . We postulated that if different RegHsa elements are predicted to regulate the same target gene, they might share common sequence motifs recognized

by the same transcription factor (TF). Consistent with this, we found significantly enriched motifs in elements targeting 124 genes (Methods), with up to 15 motifs per set of RegHsa targeting the same gene. Remarkably, different genes appear to be regulated by RegHsa elements that share the same motifs, despite the analysis being restricted to one human chromosome. The most striking case is a motif resembling the recognition sequence for the *NEUROD2* TF, present from 5 to 30 times in RegHsa elements targeting nine genes (Fig. 4a and Supplementary Data 4). *NEUROD2* is expressed in the developing brain and is important for lineage progression through chromatin remodelling<sup>14,15</sup>. Notably, several of the nine genes that are suggested here to be regulated by *NEUROD2* through common binding motifs are known to participate in different aspects of brain development and activity. In addition, 19 pairs of X-chromosome genes are linked to different sets of RegHsa elements that share three or more overrepresented motifs in common. For example RegHsa elements linked to *AFF2* and *IL1RAPL1* share five motifs in common (Fig. 4b), including a motif similar to that of the *KLF12* transcription factor, which is differentially expressed in a cellular model of neural progenitors<sup>16</sup>. Similarly, RegHsa elements linked to *BCOR* and *MAGEB10* share four overrepresented motifs (Fig. 4c) suggesting that each pair is co-regulated.

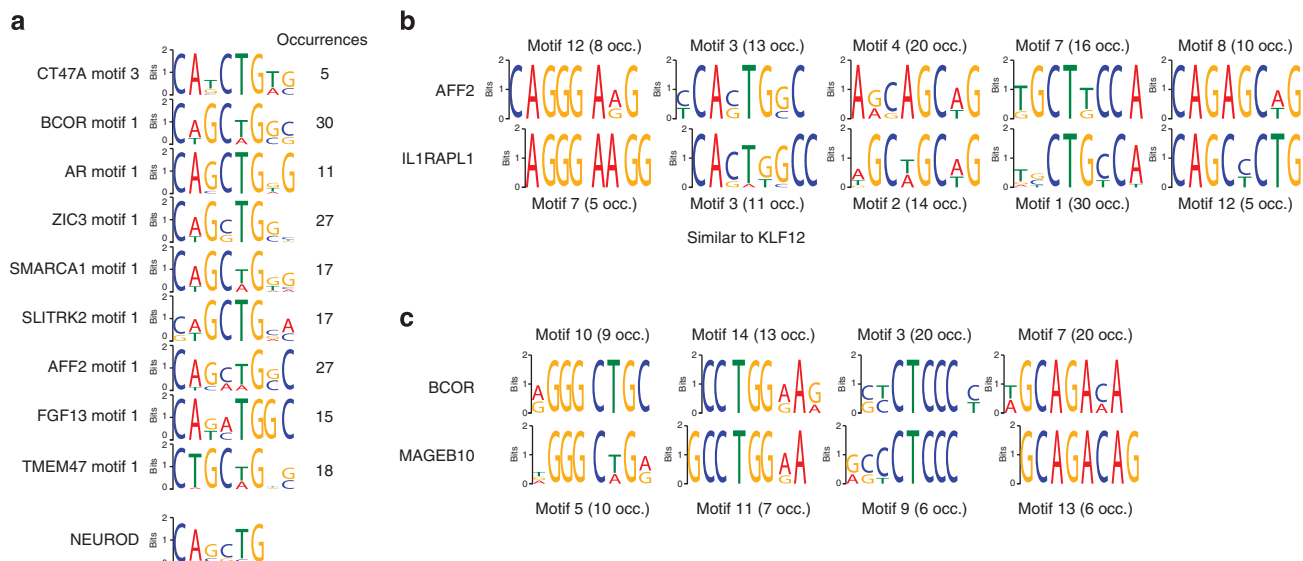
## Discussion

In summary, we describe a method to identify the evolutionary linkage between human CNEs (here, RegHsa elements) and neighbouring protein-coding target genes. We show that this linkage is indicative of a regulatory action of the element on the expression of the linked protein-coding gene. Some of these interactions were confirmed experimentally, but detailed





**Figure 3 | Neuroanatomical characterization of the element NX\_hs54.** This element includes RegHsa0032185 and was characterized in transgenic adult and juvenile zebrafish. **(a–d)** Immunohistochemical analysis of S100 $\beta$  (grey, radial glial stem cells), GFP (green), and Hu (magenta, neurons) expression in the telencephalon (level in **g**) in two different transgene integrations (2-1 and 4-1). Radial glial stem cells outline the telencephalic surface (yellow arrows, **b**) and generate neurons (white arrows, **b**)<sup>40</sup>. In one integration, GFP is expressed by virtually all neurons and their fibres underneath the radial glial cell layer (**b**). In the other integration (**c,d**), likely due to positional effects, GFP expression is restricted to individual neuronal clones (grey arrows). **(e)** *in situ* hybridization for endogenous *bcor* mRNA in the adult zebrafish telencephalon (level in **g**). *bcor* mRNA is expressed by the newborn neurons (white arrow, **f**) underlying the first cell layer of radial glial stem cells (yellow arrow, **f**). The extended GFP expression in transgenic lines is in agreement with GFP protein stability in neurons after endogenous *bcor* expression is switched off, and/or with the absence of a repressor element. **(g)** schematic lateral and dorsal views of an adult zebrafish brain showing the region (red line) examined in **a,c,e**. **(h–m)** Immunohistochemical characterization of juvenile GFP expression in NX\_hs54#4-1 demonstrates overlap with endogenous *bcor* expression (**l,m**). Use of two anatomical markers: acetylated tubulin (**h,i,j,k**; magenta) and nuclear staining (**i,k**; greyscale) permits describing GFP expression in the telencephalon at two different section levels by confocal microscopy (**h** anterior to **j**). At 3dpf in NX\_hs54#4-1 transgenic embryos GFP is widely expressed at a low level but also shows strong expression in the dorsal and lateral area adjacent to the ventricle (**h,j**; white arrowheads). This is similar to endogenous *bcor* mRNA, which also shows low level expression throughout the telencephalon and whole brain but has an area of strong expression next to the ventricle (**l,m**; yellow arrowheads, ventricle boundary marked by red dashed line). Abbreviations: AC, anterior commissure, tel, telencephalon, OB, olfactory bulb. Scale bars, **a,c,e**, 100  $\mu$ m; **b,f**, 60  $\mu$ m; **d**, 40  $\mu$ m; **h,i,j,k**, 100  $\mu$ m; **l,m**, 40  $\mu$ m.



**Figure 4 | Motifs shared between RegHsa elements suggest co-regulated genes.** (a) The NEUROD1/NEUROD2 binding site is recurrently found in multiple RegHsa elements linked to nine genes on the human X chromosome. (b) AFF2 and IL1RAPL1 share five overrepresented motifs in their linked RegHsa elements. Each motif logo is indicated together with the number of occurrences (occ.) in the set of RegHsa elements. Motif 3 is similar to the binding site of the KLF12 transcription factor. (c) BCOR and MAGEB10 share four overrepresented motifs in their linked RegHsa elements.

characterization of the different CNEs is still required. Experimental methods are already able to indicate the interactions between enhancers and genes<sup>8,13,17–19</sup> but they are strongly constrained by the tissue and time where and when the interaction takes place. In contrast, evolutionary linkage is independent of the tissue or time of expression of the gene, and is applicable to any sequenced vertebrate genome, as it was done here for human.

Regulatory mutations are known to cause diseases but few have been identified so far<sup>20–23</sup>, largely because the functional link between enhancers and their target gene is difficult to ascertain<sup>24</sup>. Here we provide a direct and simple approach to predict such interactions. For example, of the 45,449 RegHsa elements associated to one or more genes with a strong score ( $S_A > 0.8$ ), 8,217 elements target a gene where coding mutations have already been shown to cause intellectual disabilities. This strategy thus provides new material to accelerate the discovery of disease causing mutations.

## Methods

**Identification of CNEs.** CNEs are defined based on their conservation in a range of vertebrate species, using an in-house algorithm called ‘ScanMaf’ implemented in a python script (Supplementary Fig. 1a). ScanMaf scans the UCSC 46-species multiZ alignment and looks for conserved regions of a minimal length and identity, excluding exons annotated in Ensembl as well as repeats annotated by RepeatMasker and Tandem Repeat Finder. This algorithm does not require the presence of a fixed set of species in the alignment, but instead only requires a minimal number of seven species in addition to human, with no consideration of their respective phylogenetic group, allowing us to retrieve with the same procedure elements restricted to mammals as well as elements conserved between mammals and fish. It allows substitutions to occur, under a threshold of 12%, in each column of the alignment (in the minimal situation where only seven species are aligned to human, this threshold allows for one substitution); above this threshold columns are considered as conserved. The algorithm first identifies core windows of 10 bp containing at least 90% of such conserved columns. It then extends this nucleus in both the directions by allowing up to three non-conserved consecutive columns. If these human regions are conserved in the same subset of species, consecutive in each of their genomes, and separated by < 100 bp in human, they were fused in a single resulting element in order to ease further analysis. These predictions were then fused with the regions obtained by the Siphy algorithm<sup>7</sup>. The resulting 174,473 distinct CNEs on the human X chromosome were used for further analysis. Each CNE was annotated with a score to characterize its evolutionary conservation between the human sequence and the other vertebrate sequences that align to this

sequence. For this purpose, vertebrate genome sequences from the UCSC 46 species multiple alignments were classified into five groups according to their phylogenetic position: Boreoeutheria, Atlantogenata, Monotremes and Marsupials, Sauropsids and Amphibians, Teleostean fish. The maximum % ID between the human sequence and the sequences of each group, when present, are identified and summed to compute the conservation score. For example, a CNE is identified and is conserved from human to fish. The maximum % ID in each group are: Boreoeutheria 97% (with chimpanzee), Atlantogenata 68% (with elephant), Monotreme and Marsupials 62% (with opossum), Sauropsids and Amphibians 54% (with chicken) and Teleosts 49% (with medaka). The conservation score for this CNE will thus be: score = 97 + 68 + 62 + 54 + 49 = 330.

**Scoring CNE-target genes evolutionary linkage.** Families of orthologous genes were retrieved from the Ensembl database<sup>25</sup> (version 66). Starting from the human genome as a reference (version hg19), the first step of the target prediction consists in collecting immediate neighbouring genes (distant from < 1 Mb) of each given CNE within the human genome. A scoring procedure is then applied on these genes to try to identify the most probable CNE target. For any given CNE  $i$  present in  $N$  species, the absolute linkage score  $S_{Ai}$  is computed as follows:

$$S_{Ai} = \sum_1^N S_{i,e,1} \times R_e - \frac{S_{i,e,2} + C_e \times (S_{i,e,3} + S_{i,e,0})}{R_e} \quad (1)$$

where  $C_e$  is a corrective factor to minimize the influence of genome assemblies obtained at low sequence coverage (Supplementary Table 2),  $R_e$  the rearrangement rate of the genome of species  $e$  by comparison with the human genome (see below and Supplementary Table 2) and  $S_{i,e,0}$ ,  $S_{i,e,1}$ ,  $S_{i,e,2}$ ,  $S_{i,e,3}$  the respective status of the orthologous gene considered in species  $e$ :  $S_{i,e,0}$  if absent (or mis-annotated),  $S_{i,e,1}$  if present and within distance  $d$  from the CNE,  $S_{i,e,2}$  if present and beyond distance  $d$  from the CNE,  $S_{i,e,3}$  if present but on another chromosome or scaffold. These  $S_{i,e}$  parameters take the value of 1 if the condition is fulfilled, 0 otherwise. Genome coverage, rearrangement rates and distance thresholds are listed in Supplementary Table 2. Distance  $d$  is taken as 1 Mb adjusted for the size of the genome of species  $e$  compared with the human (if the genome of  $e$  is 80% of the human genome, then  $d = 0.8$  Mb). The level of synteny  $R_e$  is computed as follows:

$$R_e = \log\left(\frac{100 \times H}{P_e}\right) \quad (2)$$

where  $H$  is the total number of gene pairs in the human genome, and  $P_e$  the number of these gene pairs that are direct neighbours (in conserved synteny) in species  $e$  with the human gene pairs.  $R_e$  thus varies between 0 (a genome with no gene pairs in conserved synteny) and 1 (the human genome against itself). Of note, the baboon (papHam1) and the lamprey (petMar1) genome sequences, despite being present in the 46-species multiple alignment, were not used for the target search because of the high degree of fragmentation of their assemblies. These linkage scores, after being calculated for every gene families neighbouring each CNEs, are then normalized in a [0,1] interval using a sigmoid transformation as



follows:

$$\text{Standardization : } X = \left( \frac{x - \hat{x}}{sd(x)} \right) \quad (3)$$

$$\text{Transformation : } X_{[-1:1]} = \frac{X}{\sqrt{1+X^2}} \quad (4)$$

$$\text{Translation : } X_{[0:1]} = \frac{X_{[-1:1]} + 1}{2} \quad (5)$$

After sorting linked genes by descending linkage score, a relative score can be computed for each, corresponding to the linkage score difference between the top-ranking linkage score and the second-best linkage score. The greater the relative linkage score, the more contrast. However, if a CNE presents only one putative target in its environment, the corresponding gene family will have no relative score attributed. The relative score is useful to identify cases where, among all possible targets within 1 Mb of a given CNE, one gene stands out: this gene will have a high relative score, because there will be a high difference between its linkage score and that of the next-best target. CNEs targeting the same genes and located < 100 bp apart were fused, resulting in 102,647 RegHsa elements. The complete set of RegHsa elements together with their scores and target genes are available in Supplementary Data 1. RegHsa elements linked to their target gene with a score > 0.9 are available in a graphical interactive server on <http://www.genomicus.biologie.ens.fr/genomicus>.

**Enrichment in enhancer functional data.** Functional information was collected from the Ensembl project for DNase hypersensitive sites (DHS)<sup>26</sup>, chromatin immunoprecipitation sequencing (ChIP-seq) for TFs<sup>26</sup>, and H3K4me1, H3K4me3 and H3K27ac histones modifications<sup>26</sup> for seven different cell lines (Gm12878, H1-hESC, HSMM, HUVEC, K562, NHEK, NHLF, HMEC and NH-A). We also collected published p300 functional annotations for mouse developing heart<sup>9</sup> and mouse developing forebrain, midbrain and limb<sup>10</sup> (see below for links to public data sources). Finally, we generated p300, H3K4me1 and H3K27ac annotations using ChIP-on-chip on the human X chromosome with chromatin isolated from human fetal brain and E14.5 and P0 developing mouse brain (Methods). To compute the intersection between the functional data listed above and CNE intervals, the positions of the functional annotations and of the CNEs were compared. When the intervals overlapped by at least 1 bp, the CNE was assigned a 'functional score' corresponding to the value of the overlapping signal weighted by the percentage of the CNE covered by the signal. For instance, if a 100-bp CNE overlaps a DHS peak of value 12 over 40 bp, the DHS value associated to the CNE is:  $12 \times (40/100) = 4.8$ . For CNEs overlapping several distinct peaks, the resulting signal value is additive. In Fig. 1b and Supplementary Fig. 5, the proportion of RegHsa elements that overlap a functional annotation (with value > 0) through at least one of their constitutive CNEs was computed for each of the annotations, for classes of RegHsa elements of increasing linkage score. To associate GO<sup>27</sup> terms with X-chromosome genes predicted to be functionally linked to RegHsa elements, we used the PathwayStudio platform (Elsevier B.V., Amsterdam). GO annotations from lists of genes linked to CNEs above a certain linkage score thresholds were compared with the lists drawn from the complete list of genes of the X chromosome (Supplementary Table 1). Statistical significance was estimated by Fisher's test, without correction for multiple testing.

**Sources of public data for enhancer enrichment tests.** CNEs were annotated with a range of functional annotations, both published and obtained in the course of this project:

ENCODE (Feb. 2012) DHS<sup>26</sup>.  
(<http://genome-euro.ucsc.edu/cgi-bin/hgTrackUi?hgsid=195751083&c=chr21&g=wgEncodeAwgDnaseUniform>)  
ENCODE (Feb. 2012) ChIP-seq for Transcription Factors<sup>26</sup>.  
(<http://genome-euro.ucsc.edu/cgi-bin/hgTrackUi?hgsid=195751457&c=chr21&g=wgEncodeAwgTfbsUniform>)  
ENCODE (Feb. 2012) H3K4me1, H3K4me3 and H3K27ac histones modifications<sup>26</sup>.  
(<http://genome-euro.ucsc.edu/cgi-bin/hgTrackUi?hgsid=195751457&c=chr21&g=wgEncodeBroadHistone>).

In the 3 ENCODE data sets above, peaks correspond to local maxima of the different signals. We used data obtained in seven different cell lines (Gm12878, H1-hESC, HSMM, HUVEC, K562, NHEK, NHLF, HMEC and NH-A), by computing the mean of each functional signal in 25-bp windows along the X chromosome before intersecting these annotations with the CNE intervals.

Blow *et al.*<sup>9</sup>: p300 ChIP-seq data from mouse developing heart  
(<http://www.nature.com/ng/journal/v42/n9/xtref/ng.650-S2.xls>)

Visel *et al.*<sup>10</sup>: p300 ChIP-seq data from mouse developing forebrain, midbrain and limb.

(<http://www.nature.com/nature/journal/v457/n7231/xtref/nature07730-s2.xls>)

**Overlap between this study and interactions shown by ChIA-PET.** The genomic positions of RegHsa elements were compared with the regions shown by Li

*et al.*<sup>13</sup> to interact with gene promoters via ChIA-PET experiments. The best scoring genes of each overlapping RegHsa elements were compared with the genes that interact with the corresponding region by ChIA-PET. The two 'experiments' (linkage score in this study and ChIA-PET by Li *et al.*<sup>13</sup>) were considered consistent if one of the linked genes (with maximal score) was the same as one of the gene shown to interact by ChIA-PET. Of the 102,647 RegHsa elements identified on the human X chromosome, 2,096 elements overlap regions shown in the ChIA-PET experiment to interact with a promoter. We compared the genes evolutionarily linked (with a maximum score) with these elements, and the gene(s) shown to interact, via their promoter, with the overlapping regions via ChIA-PET. For 1,454 elements (69%), the linked genes and the interacting gene are consistent. To compute a *P* value expressing the probability of obtaining the same result by chance, we performed 10,000 resamplings of the genes linked to the 2,096 RegHsa elements that overlap ChIA-PET enhancers. In each resampling, each RegHsa element was associated to the same number of best scoring linked genes, but randomly selected among all genes present in a 2-Mb window centred on the RegHsa element. If the ChIA-PET gene target was found among these randomly associated genes, we considered the two experiments to be consistent by chance. No resampling trial reached the number of coincidences between ChIA-PET and 'linkage score' experiment obtained from in the real data. We thus estimate that the *P* value of the test is < 10<sup>-5</sup>.

**ChIP-on-chip from human and mouse developing brain.** This assay was performed for p300, H3K4me1 and H3K27ac as described<sup>28</sup>, with several modifications. Embryonic brain was isolated from human (three samples at 50 days of gestation) and mouse (E 14.5 and P0) embryos. Human fetal brain tissues were collected with informed written consent and ethical approval by Southampton and South West Hants LREC. Pools of whole brain were treated with 1.5% formaldehyde for 10 min at room temperature. Crosslinking was stopped by the addition of glycine to a final concentration of 0.125 M. The brain tissue was chopped into small pieces (~ 1 mm<sup>3</sup>) with a razor blade in cold 1 × PBS and single cell suspension was made using dounce homogenizer. The cells were swelled on ice for 10 min. in 25 mM HEPES, pH 7.8, 1.5 mM MgCl<sub>2</sub>, 10 mM KCl, 0.1% NP-40, 1 mM DTT (dithiothreitol) and protease inhibitor cocktail (Roche) and the nuclei were collected by centrifugation at 2,500 r.p.m. Nuclei were resuspended in 'sonication buffer' containing 50 mM HEPES pH 7.9, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Na-deoxycholate, 0.1% SDS and protease inhibitors, and sonicated on ice to an average length of 200–500 bp. The samples were centrifuged at 14,000 r.p.m. and the chromatin was precleared with protein-A-Dynabeads. Precleared chromatin were immunoprecipitated with 5 µg of H3K4me1 (ab8895, Abcam), 5 µg of H3K27ac (ab4729, Abcam) and 10 µg of p300 (C-20:sc585, Santacruz) antibodies and the immune complexes were collected by incubating with protein-A-Dynabeads. The beads were washed twice with 'sonication buffer', twice with sonication buffer containing 500 mM NaCl, twice with 20 mM Tris, pH 8.0, 1 mM EDTA, 250 mM LiCl, 0.5% NP-40, 0.5% Na-deoxycholate and twice with TE buffer. The immunocomplexes were eluted with 50 mM Tris, pH 8.0, 1 mM EDTA and 1% SDS at 65 °C for 10 min., adjusted to 200 mM NaCl and incubated at 65 °C overnight to reverse the cross-links. After successive treatments with 10 µg ml<sup>-1</sup> Rnase A and 20 µg ml<sup>-1</sup> proteinase-K, the samples were eluted into 50 µl H<sub>2</sub>O using the QIAquick Spin Gel Purification Kit (Qiagen). ChIP DNA and input DNA were labelled with Cy5 or Cy3, respectively, using random priming with dye-labelled random hexamers and hybridized according to the manufacturer's protocol to a HX1 (2.16 million probes) custom microarray containing specific tiled regions encompassing 99.2 and 93.8 Mb of the human and mouse X chromosome, respectively, (Nimblegen). Arrays were scanned on a NimbleGen MS 200 Microarray scanner (Nimblegen) using a laser power of 100% and 2-µm resolution and TIFF images analysed using MS 200 Data Collection software to quantitate raw signal intensities. Computational analysis of the data was carried out using the Ringo R/Bioconductor package<sup>29</sup>. The Cy5/Cy3 log<sub>2</sub> ratio were calculated for each probe and scaled by subtracting Tukey's biweight mean, as recommended in the standard manufacturer's procedure (Nimblegen). Before calling ChIP-enriched regions, we performed a smoothing over individual probe intensities. ChIP-enriched regions were called using the findChersOnSmoothed function from the Ringo package, using parameters distCutOff = 100 and minProbesInRow = 6. ChIP-chip data have been deposited to the GEO repository under accession number GSE57358. Human fetal tissue was obtained with informed consent and according to the protocol ethically approved by Southampton and South West Hants LREC. The principal investigator of these ethical approvals is D.I.W.

**Zebrafish transgenic assays of human REG elements.** Sequences chosen for testing were PCR-amplified from human genomic DNA as elements of 1–3 kb size and subcloned into pCR8 plasmid to create an entry vector for the Gateway system. Subsequent cloning into a Tol2-GFP-destination vector, microinjection of the plasmid into fertilized zebrafish eggs as well as fluorescent screening of the embryos, establishing transgenic lines and expression pattern documentation have been described elsewhere<sup>30</sup>. All the experiments were approved by the animal ethics committee of the University of Sydney and in accordance with the German protection standards and were approved by the Government of Baden-Württemberg Regierungspräsident Karlsruhe, Germany



**CNE-target gene predictions and transgenic experiments.** Transgenic elements tested in the course of this study were chosen based on a number of criteria, including sequence conservation, location near genes of medical interest and published information on enhancer function. Importantly, they were never chosen based on the linkage score described in the Methods section 2. It is therefore possible to use the transgenic experiments as a means to provide an indirect support for the two predictions:

1. The regulatory potential of the CNE, if the latter drives specific and reproducible expression of the reporter gene (GFP) during zebrafish development.
2. The target gene being regulated by the CNE, if the GFP expression pattern overlaps the expression pattern of the predicted target.

The experiment may fail to deliver an interpretable result independently of the absence of function of the CNE as human regulatory enhancer. For example, this may happen if the CNE regulates the expression of its target genes exclusively after zebrafish development is complete, if the reporter cassette (see Methods section 6) is integrated in repressive chromatin environment, or if the human sequence element is not recognized by the zebrafish orthologue of the human TF (for example, if the zebrafish ortholog has an affinity for a different sequence, or if it is altogether absent from the zebrafish genome).

Here 436 human sequence elements were tested using zebrafish transgenic experiments (Methods). These sequence elements include 1,013 RegHsa elements (Supplementary Data 2). Thereafter, results will be described and discussed in terms of RegHsa elements, because RegHsa elements are the basic 'units' of human sequences that are linked to target genes using the  $S_A$  score described in Methods. Of the 1,013 RegHsa, 574 (57%) overlap sequences that produced inconsistent expression patterns in the different F1 lines or no expression at all. The remaining 448 RegHsa produced partially or fully consistent GFP expression patterns and were further exploited. Of these, 125 elements are evolutionarily linked to one or several human genes with orthologues in zebrafish that have no recorded expression pattern in the ZFIN database. Therefore, these elements are not useful to assess the prediction that the RegHsa element is an enhancer that regulates its linked gene(s). Only the remaining 323 RegHsa elements fulfil the two conditions required to test if the transgenic experiment supports the prediction: they are contained in a sequence element that drives a partially reproducible or reproducible GFP pattern during zebrafish development, and their predicted human gene target(s) include at least one human gene with a zebrafish orthologue of known expression pattern. For the transgenic experiment, we examined the GFP expression pattern in at least five independent zebrafish F1 lines to assess the reproducibility of the pattern. The pattern was then manually recorded using ZFIN nomenclature according to the tissue(s) showing GFP expression. For the known expression pattern of the zebrafish orthologue(s), we listed the tissue(s) showing expression by *in situ* hybridization during development, or the tissue(s) affected by a mutation in the gene, or both (ZFIN database: <http://zfin.org>). The GFP expression patterns and the ZFIN expression patterns were then compared, and results show that 200 RegHsa elements (60% of 323) drive a GFP expression pattern in a tissue that is included in the published expression pattern of the predicted target, or of one of the predicted targets when several exist with an identical maximum linkage score. A schematic diagram of the decision process described here is shown in Supplementary Fig. 6. We tested the possibility that this result may be due to a bias in the RegHsa elements. Indeed, the probability of a RegHsa elements driving GFP expression by chance in at least one tissue in common with its predicted targets increases in proportion to the number of predicted targets. However, the 200 RegHsa elements that drive a GFP pattern that overlaps with the pattern of a target gene possess an average of 4.7 targets, while the 123 RegHsa elements that drive a GFP pattern that does not overlap with that of any of the target genes possess an average of 6.3 targets. Therefore the results are consistent with the starting hypothesis, that a strong evolutionary linkage score between a CNE and one or more neighbouring genes reflects a regulatory role of the CNE on the expression of one of the linked genes.

**Anatomical characterization of zebrafish GFP expression.** (a) Adult GFP expression analysis: the dissected brains of F1 adult (3–9 months) zebrafish from two different transgene integrations of each tested element were fixed in 4% paraformaldehyde for 4 h at room temperature. The following primary antibodies were applied onto free-floating 80- $\mu$ m-thick vibratome sections: GFP (1:500, chicken, Aves Laboratories), HuC/D (1:2,000, human, a gift from Dr B. Zalc, Salpêtrière Hospital, Paris), glutamine-synthase (1:500, mouse, Millipore), DAPI (diamidino-2-phenylindole; 1:3,000) was used as a nuclear counterstain. Secondary antibodies raised in goat coupled to AlexaFluor dyes (Invitrogen) were used (1:1,000). HuC/D as a neuronal marker and glutamine-synthase as a glial marker label the two main cell types of the zebrafish telencephalon<sup>31</sup> and therefore make it possible to identify GFP expressing cells. All images were taken on a Zeiss LSM700 confocal microscope using  $\times 20$  air,  $\times 40$  oil or  $\times 63$  oil objectives. Images were processed using the ZEN software (Zeiss). Composite images were automatically stitched upon acquisition using 'Tilscan' mode on the Zeiss ZEN software. (b) Adult mRNA expression analysis by chromogenic *in situ* hybridization: the dissected brains of adult (3–9 months) zebrafish from the wild-type AB strain were fixed in 4% paraformaldehyde for 14 h at 4 °C. Whole brains were incubated at 65 °C for

18 h in 2 ng  $\mu$ l<sup>-1</sup> digoxigenin (DIG)-labelled mRNA probes. After hybridization, the brains were embedded in 3% agarose and 80- $\mu$ m-thick cross sections were cut using a vibratome. The sections were blocked in blocking buffer (2% normal goat serum, 2 mg ml<sup>-1</sup> bovine serum albumin) and incubated with anti-DIG AP Fab fragments (sheep, Roche, 1:5,000) and the signal was revealed with NBT/BCIP. Pictures were taken on a Nikon AZ100 microscope equipped with a Nikon DS Ri1 camera. Expression of GFP from transgenic lines and the expression of mRNA in wild-type fish were compared manually using neuroanatomical landmarks and immunohistochemical labels. (c) Detailed expression analysis in juvenile fish: F1 juvenile zebrafish (3dpf and 6dpf) from three different transgene integrations of each tested element were anaesthetised in MS-222 and fixed immersion in 4% paraformaldehyde in 4% sucrose PBS (pH7.3). Samples were split into two sets. One set (called neuroanatomy test) was examined using wholemount immunohistochemistry to detect GFP in the context of two immunohistochemical neuroanatomical markers: SV2 and acetylated  $\alpha$  tubulin. These neuroanatomical markers provide well characterized neuroanatomical landmarks to interpret the location of GFP expression. The protocol followed was the same as that employed to prepare samples for zebrafishbrain.org<sup>32</sup>. The second set (called *in situ* test) was used to perform wholemount fluorescent *in situ* hybridization using DIG-labelled probes and tyramide detection according to the protocol of Lauter *et al.*<sup>33</sup> followed by immunohistochemical detection of GFP. Both the sets of samples were examined using confocal microscopy from a dorsal and lateral aspect (eye removed). Stacks were examined in 3D using Fiji software for neuroanatomical location and overlap between native gene expression and GFP expression. Frequently the *in situ* test set showed poor expression data for the *in situ* hybridization channel. For these sets, *in situ* hybridization was carried out on wild-type AB embryos using chromogenic detection of DIG-labelled probes according to the standard protocol of the Thisse laboratory<sup>34</sup>. Expression could then be compared between this sample and the neuroanatomical test sample. Output data took the form of text annotations of the neuroanatomical locations of GFP expression and its comparison with native zebrafish gene expression.

**Analysis of sequence motifs in RegHsa elements.** (a) *De novo* motif identification in CNEs. Conserved motifs were searched in each set of CNEs constitutive of a given REG element as long as they fulfil the following conditions (to minimize false positives): they must be associated to a single best target gene with a linkage score > 0.3 and a relative score > 0.05. Only sets comprising at least 10 CNEs (153 sets in total) were searched for possible motif enrichment. Motifs were detected using MEME3 (ref. 35) with the following options and parameters: -dna -nmotifs 15 -revcomp -mod anr -wg 6 -ws 1 -minsites 5 -maxw 8. The different motif occurrences identified by MEME in the CNEs were further reviewed to increase the motif stringency. This was done by removing sequences presenting < 80% identity with the first motif occurrence identified by MEME, which is considered to be the most similar to the motif. A threshold score characterizing each motif is then defined as the lowest weight obtained while matching the motif against each of its constitutive sequences, using the matrix-scan program of the RSATools suite<sup>36</sup>. This score will be used to seek the motif in other control CNE sets. For all RSAT tools used here, the background option ('-bgfile') was applied, with background statistics calculated on the entire set of CNEs using the oligo-analysis program with the following parameters: -l 2 -lstr -return freq. This program thus determined the frequencies of every possible dinucleotide in the total set of CNEs, and used these as background frequencies to compute the significance of observed motifs. (b) Are the motifs significantly overrepresented? Two statistical tests are further applied to eliminate motifs that may be due to chance occurrence. The first test consists in calculating a *P* value associated to the number of motif hits observed in the CNE set, by searching the motif in 1,000 random sets comprising the same number of CNEs, using matrix-scan and the weight threshold value previously computed. This *P* value reflects the number of times an equal or higher number of motif occurrences are found by chance, compared with the set of CNEs predicted to target the same gene. The second test consists in the search for motifs in the same CNE set but using shuffled motifs. These shuffled motifs are obtained by a column permutation of the motif of interest (reference motif), repeated up to 1,000 times until we obtain up to 10 motifs that are significantly different from the reference motif and from each other (the Pearson coefficient of correlation between position weight matrices, obtained by RSAT compare-matrices must be < 0.30). Motifs were ultimately considered significant with this second test if none of the shuffled matrices found > 2/3 of the number of matches found by the original motif, in the same CNE set. (c) Comparing motifs between sets of CNEs: after this filtering step, motifs obtained for distinct sets of CNEs targeting different genes were compared using the RSAT compare-matrices program<sup>36</sup>. Two motifs were considered as similar if the Pearson coefficient of correlation between their position weight matrices, further weighted by the length of the match, was > 800. (d) Are CNEs enriched in known motifs? We computed the proportion of CNEs that match known motifs, as a function of increasing evolutionary linkage score to a neighbouring gene (similar to Fig. 1b). CNEs were divided in classes of increasing linkage score, and each class was compared with the TRANSFAC database (complete vertebrate motifs; version 2010)<sup>37</sup>, to a list of sites established by high throughput SELEX<sup>38</sup> and to matrices from the JASPAR database (version 2011)<sup>39</sup> (Supplementary Fig. 4). Matches between CNEs and matrices of known motifs were identified using the matrix-scan program from RSAT<sup>36</sup>, with the background

as described above and with the following parameters: -1str -lth score 5.0. Only motifs showing a score > 15 were considered. A full description of motifs shown in Fig. 4 is in Supplementary Data 4.

**Code availability.** Python scripts to identify CNEs in multiple alignments and to compute the linkage score are freely available under a GNU GPL v3 or later, and under a CeCILL v2 license in France, as a GitHub project named Regulus: <https://github.com/DyogenIBENS/Regulus>.

## References

- Benko, S. *et al.* Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nat. Genet.* **41**, 359–364 (2009).
- Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).
- Goode, D. K., Snell, P., Smith, S. F., Cooke, J. E. & Elgar, G. Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. *Genomics* **86**, 172–181 (2005).
- Kikuta, H. *et al.* Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17**, 545–555 (2007).
- Mongin, E., Dewar, K. & Blanchette, M. Mapping association between long-range cis-regulatory regions and their target genes using synteny. *J. Comput. Biol.* **18**, 1115–1130 (2011).
- Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
- Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **7**, 476–482 (2011).
- Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
- Blow, M. J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* **42**, 806–810 (2010).
- Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
- Stevenson, R. E. & Schwartz, C. E. X-linked intellectual disability: unique vulnerability of the male genome. *Dev. Disabil. Res. Rev.* **15**, 361–368 (2009).
- Sprague, J. *et al.* The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res.* **34**, D581–D585 (2006).
- Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
- Fong, A. P. *et al.* Genetic and epigenetic determinants of neurogenesis and myogenesis. *Dev. Cell* **22**, 721–735 (2012).
- Ince-Dunn, G. *et al.* Regulation of thalamocortical patterning and synaptic maturation by NeuroD2. *Neuron* **49**, 683–695 (2006).
- Narayanan, G. *et al.* Single-cell mRNA profiling identifies progenitor subclasses in neurospheres. *Stem Cells Dev.* **21**, 3351–3362 (2012).
- Chepelev, I., Wei, G., Wangsa, D., Tang, Q. & Zhao, K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res.* **22**, 490–503 (2012).
- Merkenschlager, M. & Odom, D. T. CTCF and cohesin: linking gene regulatory elements with their targets. *Cell* **152**, 1285–1297 (2013).
- Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
- 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).
- Benko, S. *et al.* Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nat. Genet.* **41**, 359–364 (2009).
- Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum. Mol. Genet.* **21**, 3255–3263 (2012).
- Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat. Genet.* **46**, 61–64 (2014).
- Noonan, J. P. & McCallion, A. S. Genomics of long-range regulatory elements. *Annu. Rev. Genomics Hum. Genet.* **11**, 1–23 (2010).
- Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res.* **40**, D84–D90 (2012).
- Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
- Soutoglou, E. & Talianidis, I. Coordination of PIC assembly and chromatin remodeling during differentiation-induced gene activation. *Science* **295**, 1901–1904 (2002).
- Toedling, J. *et al.* Ringo—an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics* **8** (2007).
- Ishibashi, M., Mechaly, A. S., Becker, T. S. & Rinkwitz, S. Using zebrafish transgenesis to test human genomic sequences for specific enhancer activity. *Methods* **62**, 216–225 (2013).
- Marz, M. *et al.* Heterogeneity in progenitor cell subtypes in the ventricular zone of the zebrafish adult telencephalon. *Glia* **58**, 870–888 (2010).
- Turner, K. J., Bracewell, T. G. & Hawkins, T. A. Anatomical dissection of zebrafish brain development. *Methods Mol. Biol.* **1082**, 197–214 (2014).
- Lauter, G., Soll, I. & Hauptmann, G. Sensitive whole-mount fluorescent *in situ* hybridization in zebrafish using enhanced tyramide signal amplification. *Methods Mol. Biol.* **1082**, 175–185 (2014).
- Thisse, C. & Thisse, B. High-resolution *in situ* hybridization to whole-mount zebrafish embryos. *Nat. Protoc.* **3**, 59–69 (2008).
- Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
- Thomas-Chollier, M. *et al.* RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res.* **39**, W86–W91 (2011).
- Wingender, E. *et al.* The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.* **29**, 281–283 (2001).
- Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
- Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142–D147 (2014).
- Rothenaigner, I. *et al.* Clonal analysis by distinct viral vectors identifies bona fide neural stem cells in the adult zebrafish telencephalon and characterizes their division properties and fate. *Development* **138**, 1459–1469 (2011).

## Acknowledgements

This work was funded by the 7th framework programme of the European Union (NeuroXsys Project HEALTH-F4-2009-223262). T.H. and S.W. were additionally supported by BBSRC grant FBACJ 512988 and H.R.C. received support under the programme « Investissements d’Avenir » launched by the French Government and implemented by the ANR (ANR-10-LABX-54 MEMO LIFE; ANR-11-IDEX-0001-02 PSL\* Research University).

## Author contributions

M.N. and H.R.C. designed the evolutionary genomics method and M.N. performed analyses with help from A.L., M.I., M.F., E.M. S.Ri. and S.Ra. performed zebrafish transgenic experiments. H.B. performed ChIP-chip experiments. M.K. and T.A.H. performed zebrafish *in situ* experiments. C.S.R.C. analysed ChIP-chip data. D.I.W. provided human fetal tissues. D.R.F., V.V.H., S.W., B.L., U.S., L.B.-C., T.S.B., H.R.C. co-lead the project with advice from F.L.R. T.S.B. designed the initial study. M.N. and H.R.C. wrote the manuscript with contributions from L.B.-C., M.K., T.A.H., U.S., B.L., V.V.H., T.S.B. and D.R.F.

## Additional information

**Accession codes:** ChIP-chip data have been deposited to the GEO repository under accession code GSE57358

**Supplementary Information** accompanies this paper at <http://www.nature.com/naturecommunications>

**Competing financial interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available at <http://npg.nature.com/reprintsandpermissions/>

**How to cite this article:** Naville, M. *et al.* Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome. *Nat. Commun.* 6:6904 doi: 10.1038/ncomms7904 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>