Enhancer-gene maps in the human and zebrafish genomes using evolutionary linkage conservation - Supplementary information

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

1 Supplemental Text

1.1 Text S1: alignment parameters

Zebrafish-centred multiple alignments were build using LastZ [1] and Multiz [2]. Genomes used are indicated in Tables S3 & S4. We aligned zebrafish to all six species using the same parameters. Alignments parameters were as follows: step=1, masking=0, seed=12of19 (with transitions allowed), hsp threshold=3000, ydrop=3400, gapped threshold=3000, inner=0, with gap opening and extending gaps of 400 and 30. The standard HoxD55 score matrix was used for scores. The chaining step was performed on alignments with a minimum score of 0 with a loose linear gap matrix. Chained alignments were processed into nets from which best chain alignments were extracted (as indicated on the UCSC website). Pairwise alignments were merged together using Multiz. All blocks (all parameter) of minimum length 1 (R parameter) were kept.

1.2 Text S2: Long-range regulatory interactions within TADs

Topologically Associating Domains (TADs [3]) have been shown to coincide well with the regulatory landscape governing gene expression [4, 5]. Here, for CNEs linked to a single gene, 57% and 66% of predicted interactions indeed reside within a TAD in hESCs and IMR90 cells respectively [3], compared to an average of 32% and 41% respectively when we shuffle TAD intervals (proportion test p-values $< 10^{-324}$ for both cell types). CNEs are linked to their target gene with a higher score when inside the same TAD (mean scores, 0.72 inside vs 0.67 outside for hESCs, 0.71 vs 0.68 for IMR90, Wilcoxon rank sum test p-values $< 10^{-324}$ for both cell types), 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 $< 10^{-133}$) 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 $< 10^{-324}$ for both cell types).

Finally, we see a striking link between the across species conservation of CNEs and their localisation within a TAD. First, human-zebrafish orthologous CNE-target gene pairs (human-zebrafish orthologous genes with conserved CNEs) are more often located within a TAD than expected by chance (proportion tests p-values $< 10^{-30}$ for both hESCs and IMR90 cells). More importantly, we see a positive link between the conservation depth of CNEs and the co-localisation of CNEs and target genes within the same TAD. The association between TAD co-localisation and both distance to TSS and conservation depth cannot be explained by chance alone. These results, in line with previous observations [6], show evolutionary conservation of linkage between CNEs and their target genes is consistent with topological organisation of chromatin.

1.3 Text S3: Choosing a radius

PEGASUS was previously developed to predict regulatory interactions in one genome. Applying this tool to two genomes of different sizes (approximately 3 Gb for human and 1.5 Gb for zebrafish) raises the issue of the 1Mb radius to assign target genes to CNEs in both genomes. By arbitrarily setting this radius to a pre-defined value, one runs the risk of missing functional regulatory interactions located beyond this limit. We predict, however, that increasing this radius will have a negative effect on predicted interactions, as synteny conservation is more difficult to maintain over longer genomic distances. We generated CNE-target gene predictions setting the radius to a range of values between 300kbp to 2Mb in zebrafish. While the number of predicted CNEs and target genes increases linearly with the radius, the absolute unnormalized linkage score plateaus between 500kbp and 800kbp (Figure S7). We thus chose a radius of 1Mb as a compromise between the number of predicted interactions and their quality in zebrafish. In order to avoid biases in conservation of distances analyses, we chose the same radius for the human genome.

2 Supplementary Tables

hg19		***danRer7***	
anatomy term	*fe*	*anatomy term*	*fe*
endothelial cell	1.56	dorsal thalamus	8.89
lining cell	1.56	blood vessel endothelium	6.85
barrier cell	1.56	cardiovascular system endothelium	6.49
meso-epithelial cell	1.56	pretectal region	6.47
frontal pole	1.47	vestibulocochlear ganglion	5.98
pole of cerebral hemisphere	1.47	preoptic area	5.94
endothelial cell of viscerocranial mucosa	1.40	brain ventricle/choroid plexus	5.94
buccal mucosa cell	1.40	brain ventricle	5.94
cardiac muscle tissue	1.39	ventricular system of brain	5.94
myocardium of atrium	1.39	spinal cord interneuron	5.92

Table S 1: Top 10 overrepresented anatomy terms (TopAnat [7]) in human genes with conserved regulation with zebrafish. *fe*: fold enrichment. All terms have a false discovery rate lower than 0.002.

hg19		***danRer7***	
GO term	*fe*	*GO term*	*fe*
ventral spinal cord interneuron differentiation	14.02	potassium ion import	11.28
positive regulation of heart growth	12.46	central nervous system neuron differentiation	7.06
positive regulation of cardiac muscle cell proliferation	11.89	embryonic cranial skeleton morphogenesis	5.88
positive regulation of cardiac muscle tissue growth	11.60	cranial skeletal system development	5.71
central nervous system projection neuron axonogenesis	11.38	embryonic skeletal system morphogenesis	5.71
proximal/distal pattern formation	9.35	embryonic skeletal system development	5.3
positive regulation of organ growth	9.14	skeletal system morphogenesis	5.26
cell fate determination	8.90	cell fate commitment	4.88
positive regulation of cardiac muscle tissue development	8.63	skeletal system development	4.33
regulation of heart growth	8.06	positive regulation of transcription from RNA polymerase II promoter	4.19

Table S 2: Top 10 overrepresented Gene Ontology [8] terms in human genes with conserved regulation with zebrafish. *fe*: fold enrichment. All terms have a false discovery rate lower than 0.05

Common name	*Species name*	*Version*	*LCA*	*Genome size*	*control set*
Human	*Homo sapiens*	hg19	*NA*	3327	*
Chimp	*Pan troglodytes*	panTro4	HomoPan	2996	
Gorilla	*Gorilla gorilla*	gorGor3	Homininae	2829	
Orangutan	*Pongo abelii*	ponAbe2	Hominidae	3109	
Gibbon	*Nomascus leucogenys*	nomLeu3	Hominoidea	2757	
Rhesus	*Macaca mulatta*	rheMac3	Catarrhini	3094	
Marmoset	*Callithrix jacchus*	calJac3	Simiiformes	2759	
Bushbaby	*Otolemur garnettii*	otoGar3	Primates	2359	
Squirrel	*Ictidomys tridecemlineatus*	speTri2	Euarchontoglires	2311	
Mouse	*Mus musculus*	mm10	Euarchontoglires	3482	
Rat	*Rattus norvegicus*	rn5	Euarchontoglires	3042	
Guinea pig	*Cavia porcellus*	cavPor3	Euarchontoglires	2663	
Rabbit	*Oryctolagus cuniculus*	oryCun2	Euarchontoglires	2604	
Pig	*Sus scrofa*	susScr3	Boreoeutheria	3025	
Cow	*Bos taurus*	bosTau7	Boreoeutheria	2650	
Sheep	*Ovis aries*	oviAri3	Boreoeutheria	2534	
Horse	*Equus caballus*	equCab2	Boreoeutheria	2429	
Cat	*Felis catus*	felCat5	Boreoeutheria	2366	
Dog	*Canis lupus familiaris*	canFam3	Boreoeutheria	2393	
Ferret	*Mustela putorius furo*	musFur1	Boreoeutheria	2278	
Panda	*Ailuropoda melanoleuca*	ailMel1	Boreoeutheria	2245	
Microbat	*Myotis lucifugus*	myoLuc2	Boreoeutheria	1966	
Elephant	*Loxodonta africana*	loxAfr3	Eutheria	3119	
Armadillo	*Dasypus novemcinctus*	dasNov3	Eutheria	3300	
Opossum	*Monodelphis domestica*	monDom5	Theria	3502	
Tasmanian devil	*Sarcophilus harrisii*	sarHar1	Theria	2932	
Platypus	*Ornithorhynchus anatinus*	ornAna1	Mammalia	1918	
Zebra finch	*Taeniopygia guttata*	taeGut2	Amniota	1223	*
Mallard duck	*Anas platyrhynchos*	anaPla1	Amniota	1070	*
Chicken	*Gallus gallus*	galGal4	Amniota	1073	*
Turkey	*Meleagris gallopavo*	melGal1	Amniota	1062	*
Chinese softshell turtle	*Pelodiscus sinensis*	pelSin1	Amniota	2107	
Lizard	*Anolis carolinensis*	anoCar2	Amniota	1701	*
Xenopus	*Xenopus tropicalis*	xenTro7	Tetrapoda	1358	*
Coelacanth	*Latimeria chalumnae*	latCha1	Sarcopterygii	2184	

Table S 3: List of species used for PEGASUS predictions in the human genome. Species used to test the effects of phylogenetic sampling are indicated in the column "used in control set". LCA: last common ancestor with human. Genome sizes are indicated in Mb

Common name	*Species name*	*Version*	*LCA*	*Genome size*
Zebrafish	*Danio rerio*	danRer7	*NA*	1412
Medaka	*Oryzias latipes*	oryLat2	Clupeocephala	869
Tetraodon	*Tetraodon nigroviridis*	tetNig2	Clupeocephala	359
Fugu	*Takifugu rubripes*	fr3	Clupeocephala	391
Stickleback	*Gasterosteus aculeatus*	gasAcu1	Clupeocephala	462
Gar	*Lepisosteus oculatus*	lepOcu1	Neopterygii	946
Nile tilapia	*Oreochromis niloticus*	oreNil2	Clupeocephala	927

Table S 4: List of species used for PEGASUS predictions in the zebrafish genome. LCA: last common ancestor with human. Genome sizes are indicated in Mb

3 Supplementary Figures

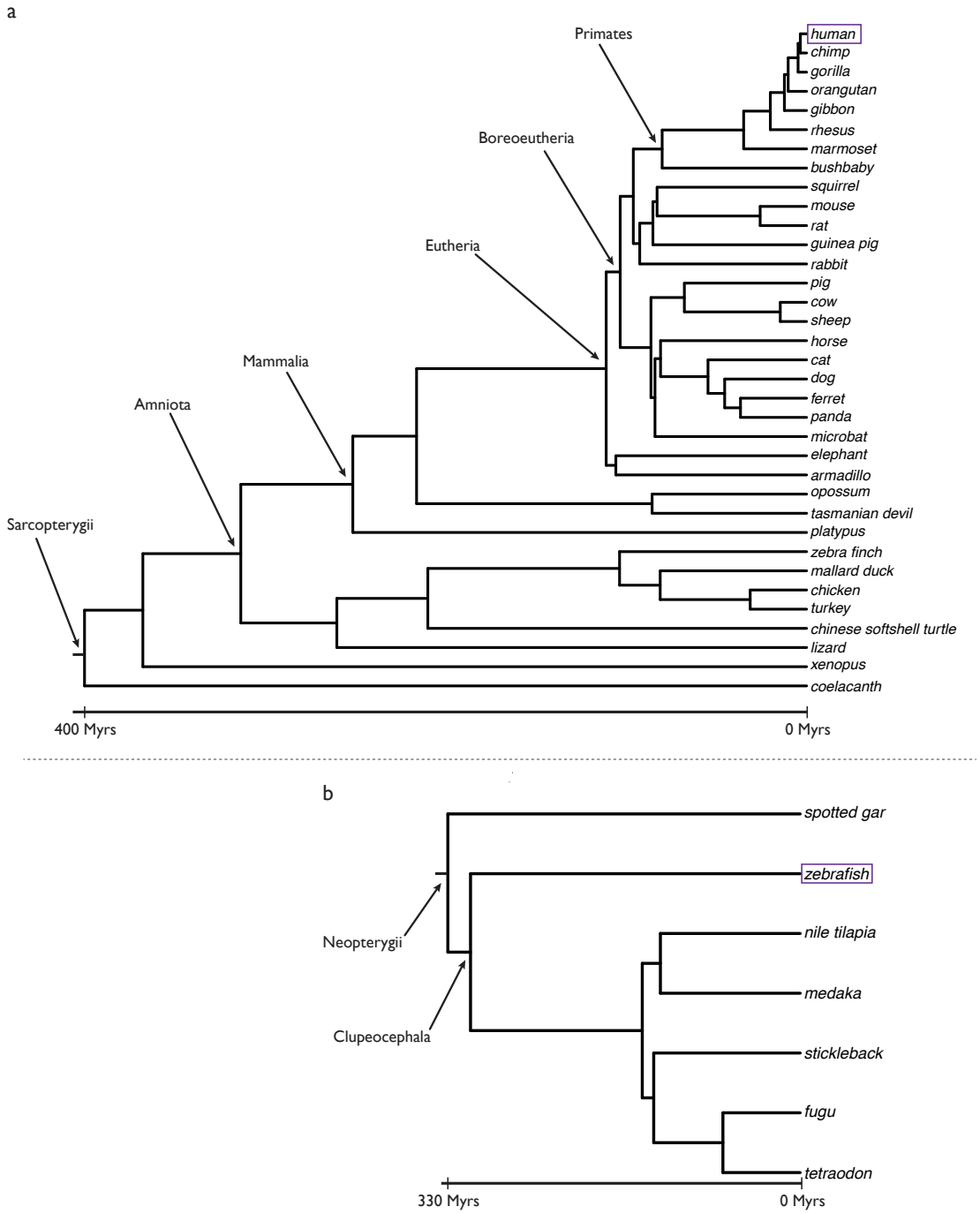

Figure S 1: Phylogenetic relationships between species used in for PEGASUS. (a) Phylogenetic relationship for the species used for the human, adapted from the tree used by the UCSC genome portal [9]. (b) The zebrafish tree was computed on a random set of 50 1-to-1 orthologous proteins using PhyML [10]. Both trees were made ultrametric using the APE package in R [11].

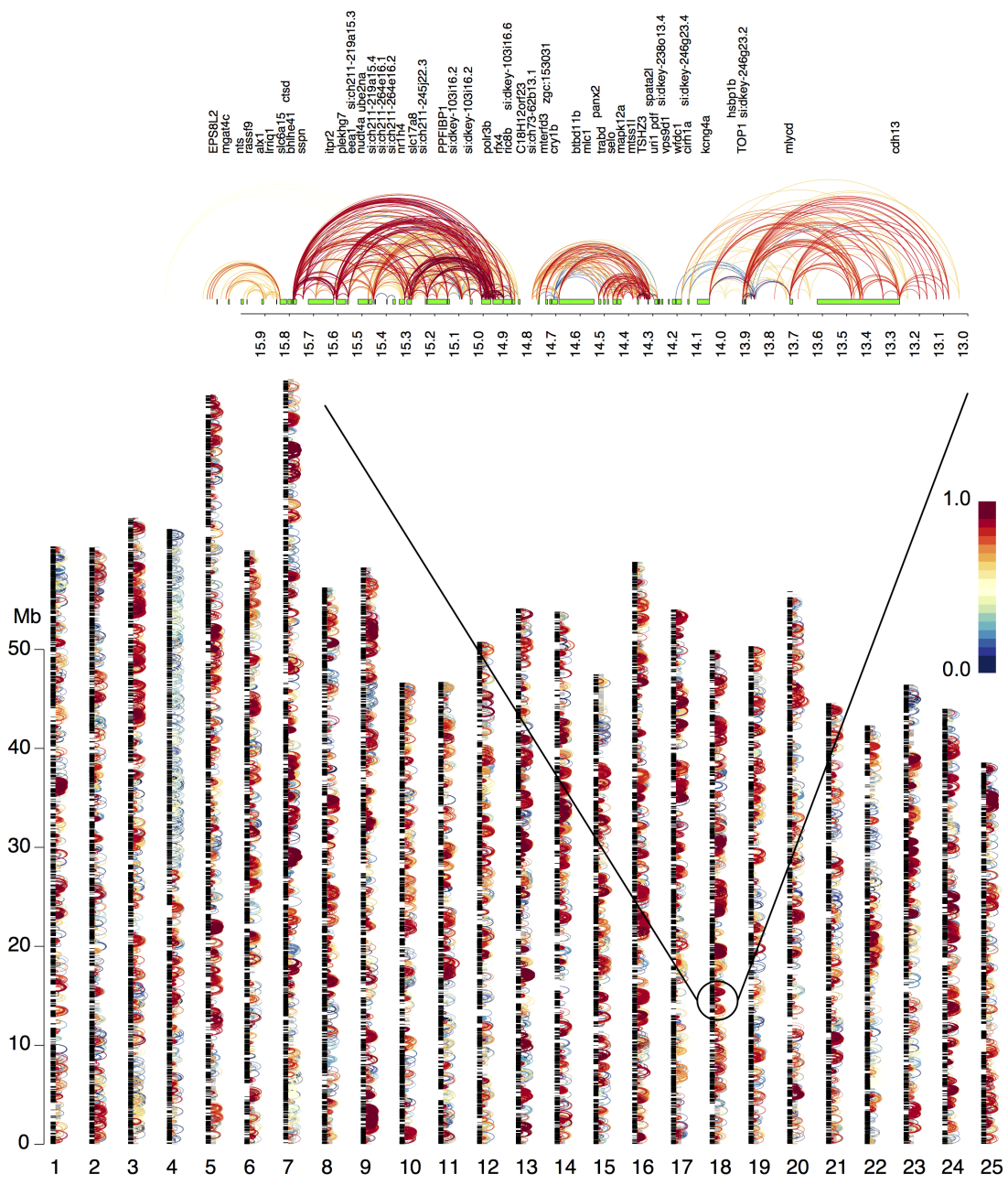

Figure S 2: Map of predicted enhancer-gene interactions in the zebrafish genome. CNE-gene associations are coloured according to their linkage scores. The top panel shows a zoom of a 3Mb region of chromosome 18.

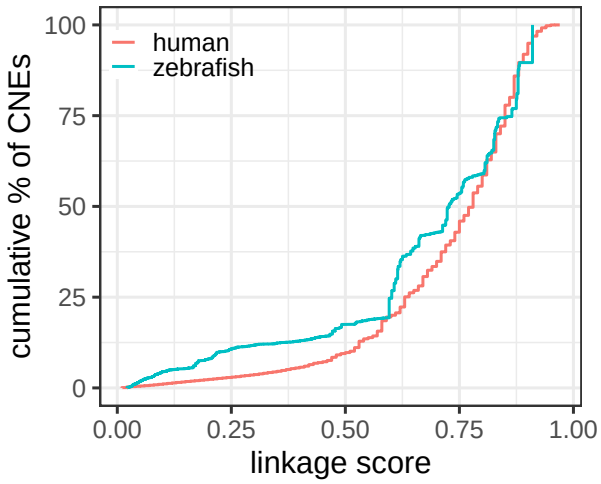

Figure S 3: Distribution of linkage scores. Cumulative distribution of linkage scores for the human and zebrafish genomes.

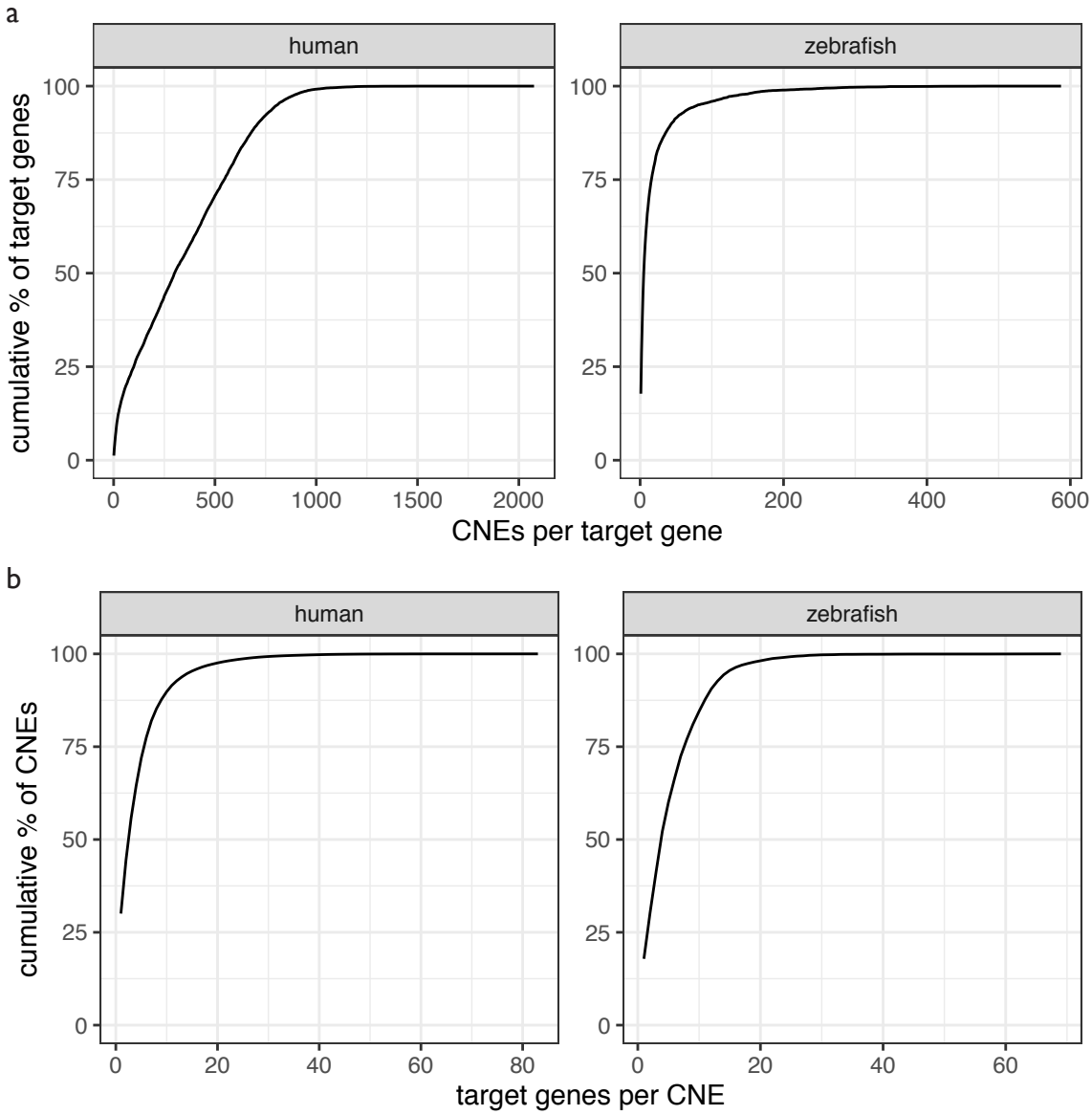

Figure S 4: Number of CNEs per target and number of target per CNEs. (a) Cumulative distribution of the number of CNEs per target gene in the human (left) and zebrafish (right) genomes. (b) Cumulative distribution of the number of target gene per CNEs in the human (left) and zebrafish (right) genomes.

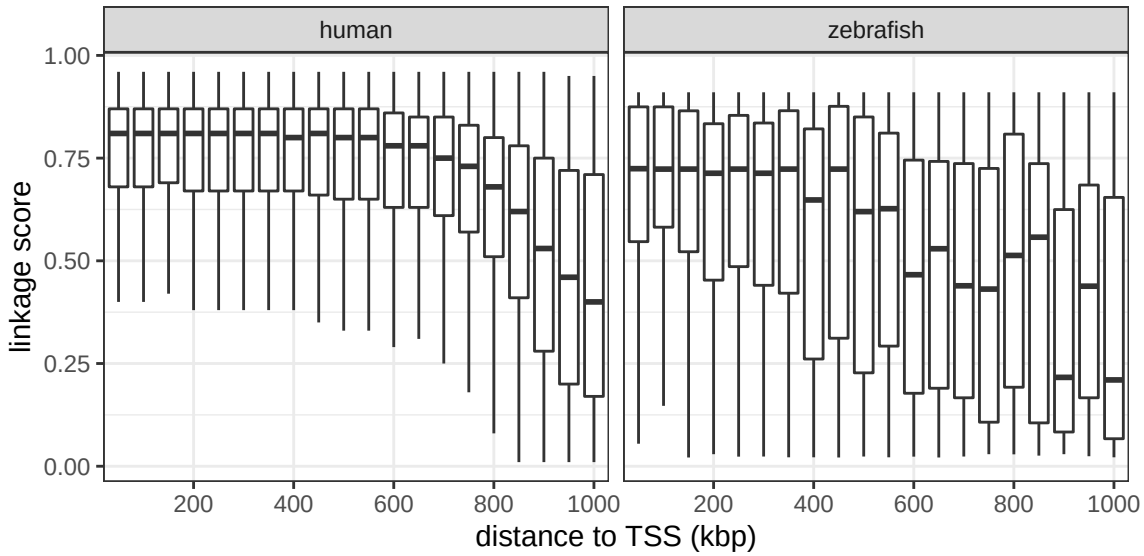

Figure S 5: Link between distance to TSS and linkage score. Histograms of CNEs' linkage score according to their distance to the TSS. Only CNEs with one target were considered.

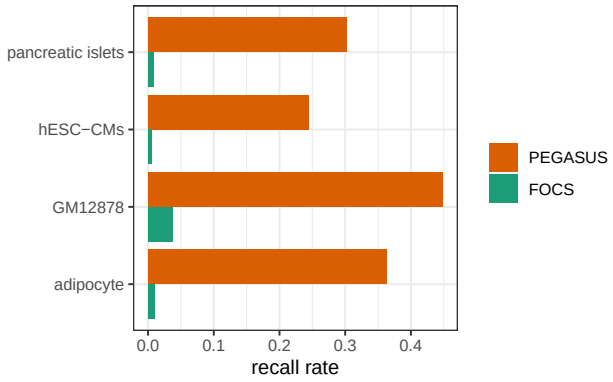

Figure S 6: Recall rates for capture Hi-C data Recall rates for four capture Hi-C datasets and FOCS (green) or the full PEGASUS dataset (orange).

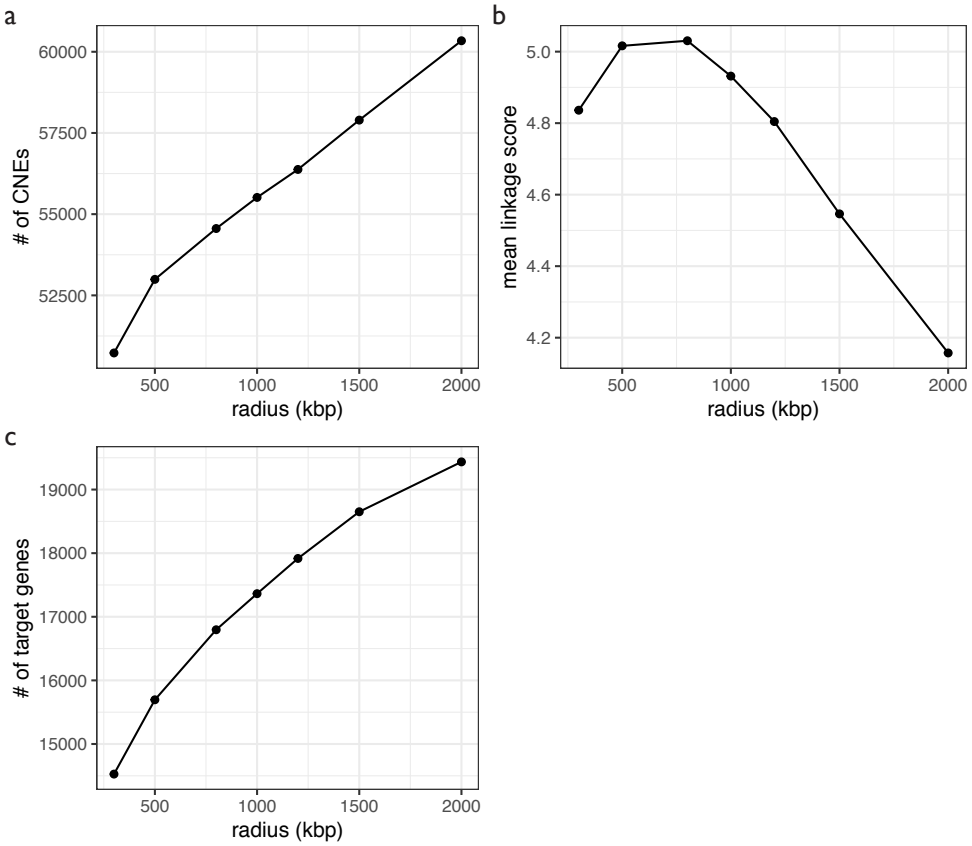

Figure S 7: Effect of radius on PEGASUS predictions (a) Number of CNEs linked to at least one target gene, (b) mean un-normalised linkage score and (c) total number of target genes as a function of the radius used for PEGASUS predictions in the zebrafish genome

References

[1] Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* (2003).

[2] Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* (2004).

[3] Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* (2012).

[4] Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* (2015).

[5] de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* (2013).

[6] Maeso, I., Acemel, R. D. & Gómez-Skarmeta, J. L. Cis-regulatory landscapes in development and evolution. *Curr. Opin. Genet. Dev.* (2017).

[7] Bastian, F. *et al.* Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. *Data Integration in the Life Sciences* (2008).

[8] Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* (2013).

[9] Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* (2002).

[10] Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* (2010).

[11] Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).