Supplemental Material

Comparative analyses of super-enhancers reveal conserved elements in vertebrate genomes

Yuvia A. Pérez-Rico[1, 2, 3], Valentina Boeva[2, 4], Allison C. Mallory[1], Angelo Bitetti[1, 3], Sara Majello[1], Emmanuel Barillot[2, 4], Alena Shkumatava[1]

[1] Institut Curie, PSL Research University, INSERM U934, CNRS UMR 3215, F-75005, Paris, France.

[2] INSERM, U900, F-75005, Paris, France.

[3] Sorbonne Universités, UPMC Univ Paris 06, F-75005, Paris, France.

[4] Institut Curie, Mines ParisTech, PSL Research University, F-75005, Paris, France.

Address correspondence to alena.shkumatava@curie.fr

SUPPLEMENTAL INFORMATION LIST

Supplemental table legends

Supplemental dataset legend

Supplemental file legends

Supplemental methods

Supplemental references

SUPPLEMENTAL TABLE LEGENDS

Supplemental Table S1. GEO sample numbers of the datasets used for typical enhancer and SE prediction in zebrafish (zf), mouse (mm) and human (hs). Total numbers of typical enhancers and SEs identified for each dataset and their median sizes.

Supplemental Table S2. Orthologous genes that have maintained their association with SEs in pluripotent state, brain, heart, intestine and testis. Human Ensembl Gene IDs are shown for the pluripotent state, brain, heart and intestine lists, and mouse Ensembl Gene IDs are shown for the testis list.

Supplemental Table S3. GEO sample numbers of the zebrafish ATAC-seq, Nanog ChIP-seq and mouse transcription factor (TF) ChIP-seq datasets. Primers used to clone zebrafish and mouse SE regions. Screening results of the enhancer reporter assays in zebrafish embryos. Matrix scanning results of the zebrafish S region defined in Fig. 6C.

SUPPLEMENTAL DATASET LEGEND

Supplemental Dataset S1. BED files with the annotated typical enhancers and SEs for each dataset.

SUPPLEMENTAL FILE LEGENDS

Supplemental File S1. Python script to identify summit coordinates.

Supplemental File S2. Bash script to filter out peaks mainly overlapping promoter regions.

25 **Supplemental File S3.** Bash script to calculate the proportion of genes covered by regions
26 of interest (e.g. typical enhancers or SEs) and perform bootstrap resampling.
27
28 **Supplemental File S4.** Python script to process hgWiggle output for one specific genomic
29 region.
30
31 **SUPPLEMENTAL METHODS**

32 **Zebrafish husbandry**

33 Wild type AB zebrafish used for this study were bred and raised following standard protocols
34 (Westerfield 2000).

35

36 **ChIP-seq assays**

37 Whole brains and testis were dissected from 10-month-old adult male AB zebrafish, whereas
38 hearts and intestines were dissected from 1-year-old adult male AB zebrafish. Two biological
39 replicates were prepared from each tissue. Each biological replicate was prepared using: 12
40 brains, 20 hearts, 5 intestines and 8 testis. All tissues were homogenized and cross-linked in
41 1% formaldehyde, washed and lysed. Chromatin was sheared using a Covaris S220
42 ultrasonicator to a DNA fragment size of 175 bp (brain and testis samples) or 200 bp (heart
43 and intestine samples). ChIP-seq was performed as previously described (Guenther et al.
44 2008) using 5 ug Abcam H3K27ac antibody (ab4729, lot# GR259887-1) bound to Dynal
45 Protein A linked beads (Invitrogen). Reverse cross-linked and phenol:chloroform purified
46 chromatin was used for single-end library preparation following standard Illumina protocols.
47 Libraries were sequenced in a HiSeq 2500 system to obtain 100 bp reads.

48

49 **Identification of typical enhancers and SEs**

50 H3K27ac ChIP-seq datasets were mapped to their corresponding reference genomes (Zv9
51 for zebrafish, mm10 for mouse and hg38 for human) using Bowtie 2 version 2.1.0
52 (Langmead and Salzberg 2012) with default parameters and allowing up to 1 mismatch in
53 the seed alignment (size = 22). Aligned reads with mapping quality ≥ 20 were saved in BAM
54 format using samtools version 1.1 (Li et al. 2009). BAM files of biological replicates were
55 merged and converted to BED format using BEDTools version 2.18 (Quinlan and Hall 2010).
56 The human brain datasets were directly downloaded from the GEO database as aligned
57 reads to the human genome version hg19 in BED format. Peak calling was performed with
58 SICER version 1.1 (Zang et al. 2009) setting window size to 200, redundancy threshold to 1,
59 gap size to 600, FDR to 0.05 and adjusting the fragment size accordingly with the analyzed
60 dataset. If available, input libraries were used as controls for the peak calling. For the
61 zebrafish datasets without input control (Supplemental Table S1) a stringent SICER e-value
62 cutoff of 0.05 was applied. Identified peaks were filtered to discard peaks for which the main
63 summit was within promoter regions (2 kb upstream and 2 kb downstream of TSSs based on
64 RefSeq annotations (Rosenbloom et al. 2015)), and if at least 50% of the peak overlapped
65 with these regions (Supplemental File S1; Supplemental File S2). To identify typical
66 enhancers and SEs the ROSE algorithm version 0.1 was applied with default parameters
67 (Whyte et al. 2013; Lovén et al. 2013) using the filtered peaks identified by SICER, H3K27ac
68 and input BAM files and performing TSS exclusion (–t 2000). H3K27ac metagene
69 representations of typical enhancers and SEs were obtained as previously described (Whyte
70 et al. 2013) by applying the “bamToGFF” function of ROSE. WIG files representing raw
71 H3K27ac occupancy were generated using FindPeaks version 4.0.15 (Fejes et al. 2008),
72 setting a triangle distribution accordingly to the fragment size of each library and removing
73 duplicates (–duplicatefilter).

74

75 **Genomic distribution of typical enhancers and SEs**

76 The calculation of typical enhancer and SE distributions around TSSs was performed using
77 the Nebula tool "Get peak distribution around TSS (histones)" (Boeva et al. 2012). Typical
78 enhancer and SE enrichments over gene bodies were calculated with a customized script
79 (Supplemental File S3) using BEDTools functions *annotate* and *shuffle*, and RefSeq Gene
80 annotations filtered to keep only unique coordinates. Enrichment over gene bodies for
81 control regions was calculated using bootstrap resampling with 100 iterations, and the mean
82 and standard deviation were used to generate Fig. 2C bar plots. To calculate the percentage
83 of typical enhancer and SE sequences overlapping with genomic features, typical enhancer
84 and SE annotations were compared to RefSeq Gene annotations (Karolchik et al. 2004;
85 Rosembloom et al. 2015) using BEDTools *intersect* function with the *–wo* option and the rest
86 as default. RefSeq overlapping regions in the same strand and for the same genomic feature
87 where merged with the BEDTools *merge* function to consider each base in the genome only
88 once. The number of base pairs overlapping with each genomic feature was divided by the
89 total length of typical enhancers or SEs to calculate the percentages.

90

91 **Gene ontology annotations**

92 Functional annotations of zebrafish SEs were performed with GREAT version 3.0.0 (McLean
93 et al. 2010; Hiller et al. 2013) using "Basal plus extension" mode and setting the distal
94 extension to 100 kb.

95

96 **Cell and tissue specificity analysis**

97 Multiple comparisons between typical enhancer and SE datasets from the same species
98 were performed with HOMER mergePeaks tool version 4.7.2 (Heinz et al. 2010) with options
99 *–d* given and *–gsize* 1412464843, 2793712140 and 3137144693 for zebrafish, mouse and
100 human, respectively. Human brain typical enhancer and SE annotations were converted

101 (liftOver; Kent et al. 2002) from hg19 genomic coordinates to hg38 coordinates for the
102 analysis.
103

104 **Sequence conservation analyses**

105 Sequence conservation scores were calculated based on the vertebrate conservation
106 PhastCons tracks from UCSC associated with each of the genome versions used for read
107 mapping (Siepel and Haussler 2005; Siepel et al. 2005). The UCSC tool hgWiggle (Kent et
108 al. 2002) was used to extract PhastCons scores for each typical enhancer and SE. A
109 customized Python script (Supplemental File S4) was written to bin typical enhancers and
110 SEs into 50 windows of equal length and process hgWiggle output file to calculate the
111 average sequence conservation of each window. Conservation values for all typical
112 enhancers and SEs and for each window were used to generate metagene representations
113 shown in Fig. 4A. Simultaneously, the average sequence conservation of the whole typical
114 enhancer or SE was calculated by the same python script. In addition, average sequence
115 conservation scores were also calculated for the immediate upstream (-3 kb) and
116 downstream (+3 kb) typical enhancer and SE regions.
117

118 **Orthologous gene comparisons**

119 Typical enhancer and SE target genes were annotated based on gene proximity using the
120 "Annotation of genes with ChIP-seq peaks (histones)" function from the Nebula web server,
121 and a maximum distance of 100 kb from gene bodies. Typical enhancer and SE annotations
122 of mouse (cerebellum and olfactory bulb only) and human adult brain and heart datasets
123 were collapsed into one single dataset for each species. Genes associated with SEs were
124 discarded from the lists of genes associated with typical enhancers. All gene names were
125 converted to Ensembl ids through bioDBnet 2.1 (Mudunuri et al. 2009) and associated with
126 the Ensembl ids of their human or mouse orthologous genes using the homology

127 annotations from Ensembl (Genes 82; Cunningham et al. 2015). For each zebrafish and
128 mouse dataset a list was generated containing the Ensembl ids of the human (pluripotent
129 state, brain, heart and intestine datasets) or mouse (zebrafish testis dataset) orthologous
130 genes and the zebrafish or mouse Ensembl ids of the genes without homology relationships.
131 The resulting lists were compared to identify shared predicted target genes between species
132 based on human Ensembl ids (for pluripotent state, brain, heart and intestine datasets) and
133 on mouse Ensembl ids (for testis datasets).

134

135 **ATAC-seq analyses**

136 ATAC-seq peaks were identified as previously described (Buenrostro et al. 2013) with
137 modifications in the programs used. Briefly, reads were mapped to the Zv9 zebrafish
138 genome using Bowtie version 0.12.8 (Langmead et al. 2009) allowing up to 1 mismatch in
139 the seed region and keeping only uniquely mapped reads (-m 1). Unaligned reads were
140 filtered to remove sequencing adaptors in the 5' region using cutadapt version 1.3 (Martin
141 2011). Trimmed reads with minimal length of 40 bp were re-mapped to the Zv9 genome. All
142 mapped reads with mapping quality ≥ 20 were merged and converted to BED format. After
143 adjusting read start sites to represent the transposon binding (Buenrostro et al. 2013), peaks
144 were identified using MACS version 1.4.2 (Zhang et al. 2008) with default band width (-bw
145 300), default range of high-confidence enrichment ratio (-m 10,30) and removing duplicates.

146 Over-represented motifs in ATAC-seq peaks within SEs were identified using the RSAT
147 *peak-motifs* tool (Thomas-Chollier et al. 2012a; Thomas-Chollier et al. 2012b). ATAC-seq
148 peaks within typical enhancers were used as background regions to perform differential
149 analysis. *Peak-motifs* was used with default parameters to identify over-represented words
150 (oligo-analysis) and spaced word pairs (dyad-analysis) using three oligomer lengths (6, 7
151 and 8) and a maximum number of 5 motifs per algorithm. *De-novo* identified motifs were
152 compared to the JASPAR core non-redundant vertebrates (2016; Mathelier et al. 2016) and

153 to the human HOCOMOCO databases (2015-11; Kulakovskiy et al. 2012). Binding sites
154 were predicted for the identified motifs using a background model with Markov order equal to
155 1. To obtain a representative set of motifs, matrix clustering was performed using the RSAT
156 *matrix-clustering* tool with an average agglomeration rule (Medina-Rivera et al. 2015).
157 Consensus motifs of the identified clusters were compared to the JASPAR core non-
158 redundant vertebrates and the human HOCOMOCO databases using *compare-matrices*
159 from RSAT with default parameters (Medina-Rivera et al. 2015).

160

161 **Zebrafish Nanog ChIP-seq analysis**

162 Reads from the Nanog ChIP-seq dataset were mapped to the zebrafish Zv9 genome version
163 using Bowtie 2 allowing up to 1 mismatch in the seed region. Reads with mapping quality ≥
164 20 were used for peak calling with MACS, using default parameters and setting the band
165 width to 150. Nanog peaks were filtered to discard peaks with FDR > 2% and sub-peaks
166 were identified using PeakSplitter version 0.1 (Salmon-Divon et al. 2010). Comparison of the
167 Nanog peaks and ATAC-seq peaks was performed with the HOMER mergePeaks tool with
168 options *–d* given and *–gsize* 1412464843.

169

170 **Analysis of mouse TF ChIP-seq datasets**

171 Mouse ChIP-seq datasets for 14 TFs were mapped to the mm10 genome version with
172 Bowtie 2 (Supplemental Table S3). Peak calling was performed with MACS, adjusting the
173 band width parameter accordingly to each library. Peaks with an FDR > 2% were discarded
174 and PeakSplitter was applied to the remaining peaks to identify sub-peaks. Sub-peaks for all
175 the libraries were collapsed and those within typical enhancers and SEs were used to create
176 enrichment tracks of TFBSs using 1 kb windows.

177

178 **Motif analysis of the SE-*zic2a* S region**

179 The SE-*zic2a* L, M, N, O and S region sequences were scanned using *matrix-scan* from
180 RSAT (Turatsinze et al. 2008) with default parameters (pseudo-counts = 1, distributed
181 proportionally to residues priors, *p*-value threshold of $1x10^{-4}$). Matrix models used for the
182 scanning corresponded to the whole sets of JASPAR core non-redundant vertebrates and
183 human HOCOMOCO matrices. Common sites between the S and the regions without
184 enhancer activity (L, M, N and O regions) were discarded and only the unique predicted sites
185 within the S region are shown in Supplemental Table S3.

186

187 **Statistical analyses**

188 All graphs and statistical analyses were performed with R version 3.1.0 (R Development
189 Core Team 2008). Venn and Chow-Ruskey diagrams were generated using the R package
190 Vennerable version 3.0 (http://r-forge.r-project.org/projects/vennerable).

191

192 **Molecular cloning**

193 SE regions were PCR amplified (Supplemental Table S3), sub-cloned into the pCRII-TOPO
194 vector (Zero Blunt TOPO PCR Cloning Kit, Invitrogen) and sequenced by Sanger method.
195 After sequence confirmation, pCRII-TOPO vectors were used to perform sticky-end
196 subcloning or Gibson assembly subcloning of the SE regions into the E1b-GFP-Tol2-
197 Gateway vector (Birnbaum et al. 2012; Addgene plasmid # 37846) with BglII following NEB
198 protocols. Ligation products were used to transform One Shot *ccdB* Survival 2 T1 Competent
199 Cells (Invitrogen). All vectors were midiprep purified (QIAGEN) and verified by sequencing.

200

201 **Microinjections of zebrafish embryos**

202 A vector carrying *Tol2* mRNA (kind gift of F. Del Bene laboratory) was linearized using NotI.
203 Capped *Tol2* mRNA was in vitro transcribed using the mMESSAGE mMACHINE SP6 Kit
204 (Life Technologies) and purified with the RNeasy Mini Kit (QIAGEN). *Tol2* mRNA was co-
205 injected with each of the E1b-GFP-Tol2-Gateway modified vectors into one-cell stage
206 zebrafish embryos. Each embryo was injected with 1 nl of a solution containing 173 ng/µl of
207 *Tol2* mRNA, 30 ng/µl of vector and 13% of phenol red. Injected embryos were kept in
208 medium containing Pen Strep and phenylthiourea (PTU) at 27°C. GFP expression was
209 monitored during the first three days post-fertilization. All injection experiments were
210 repeated at least twice (Supplementary Table S3).

211

212 **Microscopy**

213 Zebrafish embryo imaging was performed using a Zeiss SteREO Discovery.V20
214 stereomicroscope equipped with a Zeiss Axiocam MRc camera and ZEN 2012 software. All
215 images were processed using Adobe Photoshop CC software.

216

217 **SUPPLEMENTAL REFERENCES**

218 Birnbaum RY, Everman DB, Murphy KK, Gurrieri F, Schwartz CE, Ahituv N. 2012.
219 Functional characterization of tissue-specific enhancers in the DLX5/6 locus. *Hum*
220 *Mol Genet* **21**: 4930–4938.
221 Boeva V, Lermine A, Barette C, Guillouf C, Barillot E. 2012. Nebula--a web-server for
222 advanced ChIP-seq data analysis. *Bioinformatics* **28**: 2517–2519.
223 Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native
224 chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding
225 proteins and nucleosome position. *Nat Methods* **10**: 1213–1218.

226 Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham
227 P, Coates G, Fitzgerald S, et al. 2015. Ensembl 2015. *Nucleic Acids Res* **43**: D662–
228 D669.
229 Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJM. 2008. FindPeaks
230 3.1: a tool for identifying areas of enrichment from massively parallel short-read
231 sequencing technology. *Bioinformatics* **24**: 1729–1730.
232 Guenther MG, Lawton LN, Rozovskaia T, Frampton GM, Levine SS, Volkert TL, Croce CM,
233 Nakamura T, Canaani E, Young RA. 2008. Aberrant chromatin at genes encoding
234 stem cell regulators in human mixed-lineage leukemia. *Genes & Dev* **22**: 3403–3408.
235 Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H,
236 Glass CK. 2010. Simple Combinations of Lineage-Determining Transcription Factors
237 Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol*
238 *Cell* **38**: 576–589.
239 Hiller M, Agarwal S, Notwell JH, Parikh R, Guturu H, Wenger AM, Bejerano G. 2013.
240 Computational methods to detect conserved non-genic elements in phylogenetically
241 isolated genomes: application to zebrafish. *Nucleic Acids Res* **41**: e151. doi:
242 10.1093/nar/gkt557.
243 Karolchik D. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**:
244 D493–D496.
245 Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The
246 human genome browser at UCSC. *Genome Res* **12**: 996–1006.
247 Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, Makeev
248 VJ. 2012. HOCOMOCO: a comprehensive collection of human transcription factor
249 binding sites models. *Nucleic Acids Res* **41**: D195–D202.
250 Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient
251 alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25. doi:
252 10.1186/gb-2009-10-3-r25.

253 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9**:
254 357–359.
255 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin
256 R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence
257 Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
258 Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, Bradner JE, Lee TI, Young RA.
259 2013. Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers.
260 *Cell* **153**: 320–334.
261 Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing
262 reads. *EMBnet.journal* 17: 10-12. doi:http://dx.doi.org/10.14806/ej.17.1.200.
263 Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, Shi W, Shyr C, Tan G,
264 Worsley-Hunt R, et al. 2016. JASPAR 2016: a major expansion and update of the
265 open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**:
266 D110–D115.
267 McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G.
268 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat*
269 *Biotechnol* **28**: 495–501.
270 Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragón JA, Delerce J,
271 Jaeger S, Blanchet C, Vincens P, Caron C, et al. 2015. RSAT 2015: Regulatory
272 Sequence Analysis Tools. *Nucleic Acids Res* **43**: W50-W56.
273 Mudunuri U, Che A, Yi M, Stephens RM. 2009. bioDBnet: the biological database network.
274 *Bioinformatics* **25**: 555–556.
275 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
276 features. *Bioinformatics* **26**: 841–842.
277 R Development Core Team. 2008. R: A language and environment for statistical computing.
278 R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL
279 http://www.R-project.org.

280 Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR,
281 Fujita PA, Guruvadoo L, Haeussler M, et al. 2015. The UCSC Genome Browser
282 database: 2015 update. *Nucleic Acids Res* **43**: D670–D681.
283 Salmon-Divon M, Dvinge H, Tammoja K, Bertone P. 2010. PeakAnalyzer: Genome-wide
284 annotation of chromatin binding and modification loci. *BMC Bioinformatics* **11**: 415.
285 doi: 10.1186/1471-2105-11-415.
286 Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth
287 J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in
288 vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.
289 Siepel A, Haussler D. 2005. Phylogenetic hidden Markov models. In *Statistical methods in*
290 *molecular evolution* (ed. R. Nielsen), pp. 325–351. Springer, New York.
291 Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J. 2012a.
292 RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res*
293 **40**: e31. doi: 10.1093/nar/gkr1104.
294 Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J. 2012b. A
295 complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using
296 peak-motifs. *Nat Protoc* **7**: 1551–1568.
297 Turatsinze JV, Thomas-Chollier M, Defrance M, van Helden J. 2008. Using RSAT to scan
298 genome sequences for transcription factor binding sites and cis-regulatory modules.
299 *Nat Protoc* **3**: 1578–1588.
300 Westerfield M. 2000. The zebrafish book. A guide for the laboratory use of zebrafish (Danio
301 rerio), 4th Edition. University of Oregon Press, Eugene, Oregon.
302 Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young
303 RA. 2013. Master Transcription Factors and Mediator Establish Super-Enhancers at
304 Key Cell Identity Genes. *Cell* **153**: 307–319.
305 Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. 2009. A clustering approach for
306 identification of enriched domains from histone modification ChIP-Seq data.
307 *Bioinformatics* **25**: 1952–1958.

308 Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers
309 RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS).
310 *Genome Biol* **9**: R137. doi: 10.1186/gb-2008-9-9-r137.