1 CNEr: a toolkit for exploring extreme

2 noncoding conservation

3

4 Ge Tan[1,2], Dimitris Polychronopoulos[1,2] and Boris Lenhard[1,2,3]*

5 [1]Computational Regulatory Genomics Group, MRC London Institute of Medical Sciences, Du

6 Cane Road, London W12 0NN, UK

7 [2]Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, Hammersmith

8 Campus, Du Cane Road, London W12 0NN, UK

9 [3]Sars International Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgate

10 55, N-5008 Bergen, Norway

11 *Corresponding author

12

13 Ge Tan: ge.tan@fgcz.ethz.ch

14 Dimitris Polychronopoulos: dimitris.polychronopoulos@genomicsengland.co.uk

15 Boris Lenhard: b.lenhard@imperial.ac.uk

16

17 Present addresses:

18 Ge Tan, Functional Genomics Center Zurich, ETH Zurich and University of Zurich,

19 Winterthurerstrasse 190, Zurich 8057, Switzerland.

20 Dimitris Polychronopoulos, Genomics England, Charterhouse Square, London EC1M 6BQ, UK.

21

1

Abstract

Conserved Noncoding Elements (CNEs) are elements exhibiting extreme noncoding conservation in Metazoan genomes. They cluster around developmental genes and act as long-range enhancers, yet nothing that we know about their function explains the observed conservation levels. Clusters of CNEs coincide with topologically associating domains (TADs), indicating ancient origins and stability of TAD locations. This has suggested further hypotheses about the still elusive origin of CNEs, and has provided a comparative genomics-based method of estimating the position of TADs around developmentally regulated genes in genomes where chromatin conformation capture data is missing. To enable researchers in gene regulation and chromatin biology to start deciphering this phenomenon, we developed *CNEr*, a R/Bioconductor toolkit for large-scale identification of CNEs and for studying their genomic properties. We apply *CNEr* to two novel genome comparisons - fruit fly vs tsetse fly, and two sea urchin genomes - and report novel insights gained from their analysis. We also show how to reveal interesting characteristics of CNEs by coupling CNEr with existing Bioconductor packages. *CNEr* is available at Bioconductor (https://bioconductor.org/packages/CNEr/) and maintained at github (https://github.com/ge11232002/CNEr).

Introduction

44

45   Conserved Noncoding Elements (CNEs) are a pervasive class of extremely conserved elements

46   that cluster around genes with roles in development and differentiation in Metazoa [1,2]. While

47   many have been shown to act as long-range developmental enhancers [3,4], the source of their

48   extreme conservation remains unexplained [5,6]. The need to maintain arrays of CNEs in cis to

49   the genes they regulate has led to their spatial arrangement into clusters termed Genomic

50   Regulatory Blocks (GRBs) [7,8]. The role of those clusters in genome organisation is suggested

51   by recent findings demonstrating that ancient metazoan clusters of extreme noncoding

52   conservation coincide with topologically associating domains (TADs) [9].

53

54   Numerous recent studies highlight and seek to elucidate the importance of functional non-

55   coding regions, most recently by employing the CRISPR-Cas9 based techniques to locate and

56   dissect elements that affect gene expression and phenotype/disease - associated processes

57   [10–12]. Prioritizing target loci of interest for interrogating the function of their regulatory context

58   will be one of the major focuses of functional genomic studies, as has been shown in the case

59   of the *POU5F1* locus [13], and of *NF1*, *NF2* and *CUL3* genes [14]. CNEs and the regulatory

60   landscapes defined by their clusters serve as excellent candidates for such studies [3,15,16].

61

62   A handful of CNE resources exist, mainly databases, which contain already pre-computed

63   clusters of CNEs. These databases are static and mostly not updated. A summary of these

64   resources is available in the review by Polychronopoulos et al. [6]. To our knowledge, there are

65   only two tools available for the identification of conserved elements: PHAST [17] and

66   CNEFinder [18]. The former relies on multiple sequence alignments and requires extensive

67   computation time to derive "conserved" and "non-conserved" states from a two-state

68   phylogenetic hidden Markov model (phylo-HMM), a space-time probabilistic model that

69   considers both the nucleotide substitution of each base in the genome sequence through

70   evolution and the transition from one base to the next. The latter produces CNEs based on a k-

71   mer technique for computing maximal exact matches thus finding CNEs without the requirement

72   of whole-genome alignments or indices. Neither of them comes with a comprehensive, easy-to-

73   follow suite of tools tailored to the integrated exploration of CNEs from end-to-end: from

74   identification to quality control and visualisation. Our package couples those processes

75   together, enabling the user to harness the support and wealth of packages available through the

76   common the Bioconductor infrastructure. Our package is specifically designed for efficient

77   identification of CNEs using user-specified thresholds, and it functions equally well across

78   vertebrates, invertebrates or plants. To study the evolutionary dynamics of these elements and

79   their relationship to the genes around which they cluster, it is essential to be able to both

80   produce and explore genome-wide sets of CNEs for a large number of species comparisons in

81   a dedicated workflow, each with multiple length and conservation thresholds.

82

83   The *CNEr* package aims to detect CNEs and visualise them along the genome under a unified

84   framework. For performance reasons, the implementation of CNE detection and corresponding

85   I/O functions are primarily written as C extensions to R. We have used *CNEr* to produce sets of

86   CNEs by scanning pairwise whole-genome net alignments with multiple reference species, each

87   with two different window sizes and a range of minimum identity thresholds, available at

88   http://ancora.genereg.net/downloads. In this work, we demonstrate the application of *CNEr* to

89   the investigation of noncoding conservation between fruit fly *Drosophila* and tsetse fly *Glossina* -

90   the two species at the evolutionary separation not previously investigated in insects [7] - and

91   between two species of sea urchins. This has enabled us to observe some properties of GRB

92   target genes shared across Metazoa. In a previous study, we showed that more distant

93   comparisons in Diptera (between Drosophila and mosquitoes) failed to identify CNEs [7]. On the

94   other hand, the conservation level across different species of the Drosophila genus is

4

95   comparable to that across placental mammals. With *Drosophila* and *Glossina*, we wanted to

96   explore the evolutionary divergence comparable to human vs. fish in another lineage and

97   establish whether it is the same functional class of genes that is accompanied by CNEs

98   featuring such a deep level of conservation. In the case of sea urchins, we wanted to investigate

99   a lineage at an intermediate distance to vertebrates - closer than insects, more distant than the

100   early branching chordates - in order to establish the continuum of GRBs across Metazoa. We

101   present a series of downstream analysis of the newly identified CNEs, identifying their

102   characteristic sequence features in invertebrates and functional classes of genes whose loci

103   they span.

104   Design and Implementation

105   Overview of *CNEr* workflow

106   *CNEr* provides the functionality of large-scale identification and advanced visualisation of CNEs

107   based on our previous strategies of detecting CNEs [7,8,19] as shown in Fig 1. *CNEr* scans the

108   whole genome pairwise net alignment, which can be downloaded from UCSC or generated by

109   the *CNEr* pipeline, for conserved elements. Various quality controls of the alignments are

110   provided. The composition of aligned bases in the alignment can be used for tuning parameters

111   during pairwise alignment (S1 Fig). More closely related species are expected to give higher

112   rates of matched bases. The syntenic dotplot of the alignments (S2 Fig) quickly shows the

113   syntenic regions between two assemblies.

114

115   **Fig 1: *CNEr* workflow.**

116   (A) A typical pipeline of identification and visualisation of CNEs. (B) Illustration of scanning an

117   alignment for CNEs. The scanning window moves along the alignment for conserved regions.

118   The exons and repeats regions are skipped during the scanning by default.

119

120   Considering the different extents of evolutionary divergence and sequence similarity between

121   assemblies, we typically use the identity thresholds of 70% to 100% identity over a scanning

122   window of 30 bp or 50 bp. Known annotations of exons and repeats are compiled from sources

123   such as UCSC [20] and Ensembl [21] for common genomes, and elements overlapping with

124   these regions are typically skipped during the scanning. Genome annotation pipeline, such as

125   MAKER [22], can be used to create annotations for new genome assemblies.

126

127   Net alignments only keep the best match for each region in the reference genome. This is not

128   acceptable when one of the aligned genomes underwent one or more whole genome

129   duplications, leading to legitimate deviations from 1:1 orthology for many CNEs. To eliminate

130   the bias of the choice of reference genome in the alignment and to capture duplicated CNEs

131   during whole genome duplication (WGD), we scan two sets of net alignments by using each of

132   the two compared genomes as reference in turn. This strategy performs well when comparing

133   species with different numbers of WGD rounds, such as tetrapod vertebrates against teleost fish

134   [23], or common carp [24] against other teleost fish. In such cases, some of the identified CNE

135   pairs from two rounds of screening do overlap on both assemblies, and hence are merged into

136   one CNE pair. As the last step, we align the CNEs back to the two respective genomes using

137   BLAT and discard the ones with high number of hits. The remaining elements are considered to

138   be a reliable set of CNEs.

139

140   *CNEr* provides a quick overview of the genomic distribution of CNEs along chromosomes. In S3

141   Fig, each CNE between human and mouse is plotted relative to each human chromosome (x-

6

142 axis). We plot the cumulative number of CNEs over chromosomal positions. A CNE cluster is

143 represented as a sharp increase of height in y-axis with small change in x-axis. For visualisation

144 of CNEs in any genome browser, *CNEr* can export the CNE coordinates in BED file format and

145 CNE density (measured by the percentage of area covered by CNEs within a smoothing

146 window) in **bedGraph** and **bigWig** formats. Since running the whole pipeline of CNE detection

147 can be time-consuming, we also implemented a storage and query system with SQLite as

148 backend. Based on the visualisation capability of the *Gviz* package [25], *CNEr* can produce

149 publication-quality horizon plots of CNE density along with other genomic annotations (see

150 Methods and Data). Examples of the horizon plots are given in the following sections.

CNEr package implementation

152 *CNEr* is a Bioconductor package developed in R statistical environment, distributed under the

153 GPL-2 licence for *CNEr* code, and UCSC Kent's licence for Jim Kent's C source code it builds

154 on [20]. Although *CNEr* supports compilation for both 32-bit and 64-bit systems across multiple

155 platforms, it has limited functionality on the Windows platform due to the lack of the external

156 sequence alignment software *BLAT* [26], which is required in the pipeline.

Overview of whole genome pairwise alignment pipeline

158 UCSC Genome Informatics (http://hgdownload.soe.ucsc.edu/downloads.html) provides the

159 pairwise alignments between many popular species. However, there is a frequent need to

160 produce pairwise alignments for novel genome assemblies for new species, or using specific

161 assembly versions when they are not available from UCSC. This pipeline mostly requires

162 external sequence aligners and UCSC Kent's utilities [20], and provides well-tested parameters

163 for species with a varying degree of evolutionary divergence. In brief, first a sequence alignment

164 software, *LASTZ* [27] or LAST [28], is used to find the similar regions between two repeat-

165   masked genomes. Then, if two neighbouring alignments are close enough in the genome, they

166   are joined into one fragment. During the alignment, every genomic fragment can match several

167   others, and the longest one is kept. Finally, blocks of alignments are grouped into stretches of

168   synteny and form the so called "net" alignments in Axt format [29]. *CNEr* comes with a vignette

169   to demonstrate the whole pipeline. The produced Axt alignment can be manipulated in R as the

170   *Axt* class, which is extended from *GRangePairs* class defined in *CNEr* (see S2 Text).

171 Overview of the Axt scanning algorithm

172   The Axt alignment scanning algorithm constitutes the central part of this package for the

173   identification of conserved noncoding elements. Due to the massive manipulation of characters,

174   we implemented this algorithm purely in C for performance reasons; it is available to the R

175   environment through R's C interface. The minimal input is the Axt alignment and the ranges to

176   filter out, i.e., the coding and/or repeat masked regions.

177

178   The Axt screening algorithm proceeds as in S1 Algorithm. First, the Axt alignment is converted

179   into a linked Axt data structure as implemented in Jim Kent's UCSC source code [20]. The

180   filtering ranges are encoded into a hash table, where keys are the chromosome/sequences

181   names and values are pointers to the linked lists of coordinates ranges. We then iterate over the

182   linked Axt alignments. For each alignment, we use a running window to scan the alignment with

183   a step size of 1 bp. Each base is searched against the filtering hash table and matched bases

184   are skipped. All segments above the identity threshold are kept. The overlapping segments are

185   merged into larger pieces. This procedure produces a set of CNEs conserved between the two

186   aligned genome assemblies.

CNEr visualisation capability

188 Instead of using the standard density plot for CNE density (as implemented in e.g. the Ancora

189 browser), we introduce the horizon plot with the aim to increase the dynamic range of CNE

190 density visualisation. The horizon plot provides a way of visualising CNE density over several

191 orders of magnitude, and eliminates the need for multiple standard density tracks at different

192 thresholds along the genomic coordinates. Instead, a relatively low conservation threshold is

193 used, and multiple overlaid sections of the horizon plots will reveal peaks with different

194 conservation density (see Fig 3A and Fig 3B in horizon plot, Fig 3C and Fig 3D in Ancora

195 browser). We expand the functionality of "horizonplot" in *latticeExtra* package and integrate it

196 into *Gviz* [25], which is the plot engine used in *CNEr*.

Results

CNEr use case I: *Drosophila-Glossina* CNEs

199 Here we demonstrate the application of *CNEr* to the analysis of Tsetse Fly (*Glossina morsitans*)

200 CNEs and their putative target genes. *Glossina* is the sole vector of African trypanosomiasis

201 ("sleeping sickness"), and it mediates transmission of the disease during feeding on blood. It

202 has been shown previously [7] that, while there are tens of thousands of CNEs detected across

203 different *Drosophila* species, there are almost no highly conserved elements found between

204 *Drosophila* and malaria mosquito *Anopheles gambiae* or other mosquitos. *Glossina* and

205 *Drosophila* are much closer to each other than either of them is to mosquitos, having a common

206 ancestor that has diverged around 60.3 Mya (S4 Fig). With the recently available assembly and

207 gene annotation of *Glossina* [30] (see S1 Text), we were able to identify clusters of CNEs

208 between these two species. The clusters correspond to a subset of clusters defined by the

9

209   CNEs derived from comparisons of different *Drosophila* species. A further investigation of gene

210   functions, which are retained or missing in *Glossina*, was carried out by comparison with the

211   *Drosophila* clusters.

212

213   A summary of CNEs detected between *Glossina* and *Drosophila* is given in Table 1. As

214   expected, many fewer CNEs are detected from the comparison between *Glossina* and

215   *Drosophila* than between any two *Drosophila* species, since *Glossina* is an outgroup to the

216   *Drosophila*/*Sophophora* family. A closer examination of the CNE density plot in Ancora browser

217   [31] revealed many missing clusters of CNEs relative to CNE density across *Drosophila*

218   species, especially at a more stringent threshold. We wanted to find out if the missing and

219   retained CNE clusters differ with respect to the functional categories of the genes they span. In

220   the following analysis, the CNEs that are conserved for more than 70% over 30 bp are

221   considered.

222

Table 1: The number of CNEs between D. melanogaster and several other species,

including G. morsitans

Minimum identity	vs D. ananassae	vs D. pseudoobscura	vs D. mojavensis	vs D. virilis	vs G. morsitans
70% over 30 bp	NC	NC	176366	204970	9691
80% over 30 bp	NC	313570	127293	146793	3924
90% over 30 bp	NC	212951	81436	92288	1922
96% over 30 bp	177759	128843	47408	52134	813

100% over 30 bp	112073	76715	26972	29445	414
70% over 50 bp	266385	248357	104476	120628	3185
80% over 50 bp	223975	177266	66063	75204	1796
90% over 50 bp	142899	96994	33455	37098	732
96% over 50 bp	79631	49380	16387	17831	244
98% over 50 bp	55460	33463	10741	11548	150
100% over 50 bp	29218	17201	5250	5585	66

NC, not counted due to the threshold being too low for close species.

223

224 The most deeply conserved vertebrate CNEs are usually associated with genes involved in

225 transcriptional regulation or development (trans-dev) functions [19]. Due to high divergence

226 between *Drosophila* and *Glossina*, the regions with detectable CNE arrays tend to be of low

227 CNE turnover, i.e. the process of sequence divergence and loss of ancestral CNEs is slow. If

228 the same functional subset of genes is surrounded by low-turnover CNE clusters as in

229 vertebrates, the encompassed genes will more likely be essential key developmental genes [5].

230 Indeed, *Drosophila* genes associated with (i.e. nearest to) *Glossina* vs. *Drosophila* CNEs are

231 also associated with trans-dev terms (Fig 2A). Development, including organ, system and tissue

232 development, appears at the majority of the top Gene Ontology (GO) terms. The other highly

233 significant GO terms include biological regulation, regulation of cellular process and cell

234 differentiation. CNE clusters can span regions of tens to hundreds of kilobases around the

235 actual target gene, which is on average shorter than the equivalent spans in vertebrate

236 genomes. This is in agreement with our observation that CNE clusters and the GRBs they

11

237   define (and, by extension, the underlying TADs) expand and shrink roughly in proportion to

238   genome size [9]. The *H15* and *mid* locus (Fig 3A) is one of the biggest CNE clusters retained

239   between *Glossina* and *Drosophila*. The *H15* and *mid* genes encode the T-box family proteins

240   involved in heart development [32]. Although the CNE density between *Drosophila* and *Glossina*

241   is much lower than that across the *Drosophila* genus, it clearly marks the CNE cluster

242   boundaries of this locus, containing 67 CNEs at the 70% identity over 30 bp threshold. For the

243   40 largest retained CNE clusters, we provide a comprehensive list of CNE cluster coordinates,

244   the target genes, the protein domains and the number of associated CNEs (S1 Table). As we

245   can see, the majority of the target genes have *Homeobox*, *Forkhead* or *C2H2* Zn finger

246   domains, just like the genes spanned by the most conserved CNE clusters in vertebrates.

247

248   **Fig 2: Over-represented GO Biological Process terms ranked by GeneRatio.**

249   The gene ratio is defined as the number of genes associated with the term in our selected

250   genes divided by the number number of selected genes. The p-values are adjusted using "BH"

251   (Benjamini-Hochberg) correction. The visualisation is done by *clusterProfiler [33]*. (A) GO

252   enrichment for genes nearest to *Drosophila* and *Glossina* CNEs. (B) GO enrichment for genes

253   in the missing CNEs clusters compared between *Drosophila* and *Glossina*.

254

255   **Fig 3: Horizon plot of CNE density around key developmental genes along *D.***

256   ***melanogaster* as reference.**

257   (A) *H15* and *mid* genes are spanned by arrays of CNEs. Despite the much lower CNE density

258   from *D. melanogaster* and Glossina, a CNE cluster boundary shows up that is consistent with

259   CNEs from other drosophila species. (B) The CNE cluster around *ct* gene is missing in the

260   comparison of *D. melanogaster* and Glossina since no CNEs are detected. This implies that this

261   region undergoes a higher CNE turnover rate. (C, D) The same loci as in (A, B) are shown on

262   the Ancora browser in order to compare the normal CNE density plot with the horizon plot.

12

263   Notations: ensGene, Ensembl gene track; *Glossina* 21/30, G. morsitans 70% identity over 30

264   bp; droAna2 49/50, *D. ananassae* 98% identity over 50 bp; dp3 48/50, *D. pseudoobscura* 96%

265   identity over 50 bp; droMoj2 48/50, *D. mojavensis* 96 % identity over 50 bp; droVir2 48/50, *D.*

266   *virilis* 96% identity over 50 bp.

267

268   Some other regions have strong clusters of CNEs conserved among *Drosophila* species,

269   however, the CNE cluster between *Drosophila* and *Glossina* is absent. The *ct* locus (Fig 3B),

270   encoding the cut transcription factor, is one of the best representative cases. Ct plays roles in

271   the later stages of development, controlling axon guidance and branching in the development of

272   the nervous system, as well as in the specification of several organ structures such as

273   Malpighian tubules [34]. In order to locate the CNE clusters missing from *Drosophila* vs.

274   *Glossina* comparison, we use the CNE clusters detected in *D. melanogaster* vs. *D. ananassae*

275   comparison as reference and compare them with the aforementioned retained CNE clusters.

276   The genes within those missing CNE clusters are highly enriched for axon guidance and

277   neuronal development (Fig 2B). We then examine the CNE turnover rate (the speed of replacing

278   old CNEs) of the 216 human genes that are associated with the axon guidance term

279   (GO:0007411), with both human and *Drosophila* as reference. The turnover rate is calculated as

280   the reduction of the number of CNEs between two sets of CNEs. For human reference, we

281   choose the CNE set of human vs. mouse and human vs. zebrafish, while *D. melanogaster* vs.

282   *D. ananassae* and *D. melanogaster* vs. *Glossina* are chosen for *Drosophila* reference. As

283   shown in Fig 4, the axon guidance genes have significantly higher turnover rate than the other

284   genes ($p < 1e-5$, Kolmogorov-Smirnov one-sided test) in both human and *Drosophila* lineages.

285

286   **Fig 4: Cumulative distribution function of the changes of CNE number.**

287   For a 40kb window around each orthologous gene pair between human and *drosophila*, we

288   calculate the reduction of the number of CNEs for human (# of CNEs from human-mouse

13

289   comparison minus # of CNEs from human-zebrafish comparison) and drosophila (# of CNEs

290   from *D. melanogaster* vs. *D. ananassae* comparison minus # of CNEs from *D. melanogaster* vs.

291   *Glossina*) as reference. The axon guidance genes significantly show a higher degree of CNE

292   number reduction, compared with the other genes ($p < 1e-5$, Kolmogorov-Smirnov one-sided

293   test).

CNEr use case II: sea urchin CNEs

295   In this section we apply *CNEr* to the comparison of highly fragmented genome assemblies of

296   two sea urchin species *Strongylocentrotus purpuratus* and *Lytechinus variegatus* (see S1 Text).

297   The purpose of this analysis is twofold. First, we want to demonstrate how well *CNEr* is able to

298   call CNEs and their clusters in the case of highly fragmented draft genomes: the ability to

299   perform this analysis on draft genome assemblies would show that our approach can be applied

300   to a large number of available genomes, most of which haven't been assembled past the draft

301   stage and are likely to remain in that state. Second, we wanted to ask if a third lineage,

302   evolutionarily closer to vertebrates than insects but still lacking any shared CNEs with

303   vertebrates, would exhibit the same patterns of noncoding conservation. This could provide a

304   hint towards CNEs' universal presence in Metazoa, in addition to providing an informative

305   additional dataset for comparative studies of genomic regulatory blocks.

306

307   *S. purpuratus* is a popular model organism in cell and developmental biology. *S. purpuratus* and

308   *L. variegatus* have a divergence time of 50 Mya [35] and historically moderate rates of

309   sequence divergence, which makes them ideal for comparative genomics studies of regulatory

310   elements. We identified 18,025 CNEs with threshold of 100% identity over 50 bp window.

311   Despite the highly fragmented assemblies, we could clearly detect 808 prominent CNE clusters.

312

313   An especially interesting observation is the largest cluster we detected, at the *Meis* gene locus

314   (Fig 5). The CNE density clearly marks the boundaries of the CNE cluster. In Metazoa, *Meis*,

315   one of the most well-known homeobox genes, is involved in normal development and cell

316   differentiation. Tetrapod vertebrates have three *Meis* orthologs as a result of two rounds of

317   whole genome duplication. The CNE cluster around *Meis2* (one of three Meis paralogs arisen

318   by two WGD rounds at the root of vertebrates) is the largest such cluster in vertebrates [19].

319   Remarkably, the cluster of CNEs around *Drosophila*'s *Meis* ortholog, *hth* (homothorax), is also

320   the largest CNE cluster in the *D. melanogaster* genome [7]. It is currently unknown why the

321   largest clusters of deeply conserved CNEs are found around the same gene in three different

322   metazoan lineages, even though none of the CNEs from one lineage has any sequence

323   similarities to CNEs in the other two. The most plausible explanation is that the ancestral *Meis*

324   (*hth*) locus was already the largest such locus in the ancestral genome, and that CNE turnover

325   has led to three separate current lineage-specific sets of CNEs.

326

327   **Fig 5: Horizon plot of CNE density at *Meis* loci on sea urchin *Strongylocentrotus***

328   ***purpuratus.***

329   The density plots of CNEs detected at similarity threshold 96% (48/50), 98% (49/50) and 100%

330   (50/50) over 50 bp sliding window, in *Lytechinus variegatus* comparison, are shown in three

331   horizon plot tracks. The boundaries of CNE clusters from various thresholds are mutually

332   consistent.

333

CNEr reveals interesting sequence features characteristic of ultraconservation

It has been shown that vertebrate nonexonic CNEs are enriched in the TAATTA hexanucleotide motif, which is an extended recognition site for the homeodomain DNA-binding module [36]. With *CNEr*, we can easily verify the existence of TAATTA motif in CNEs of invertebrate species. In S5 Fig A, we consider CNEs identified by *CNEr* that are conserved between *D. melanogaster* and *D. virilis* over 98% for more than 50 nucleotides and plot them by increasing width using *heatmaps* package (https://bioconductor.org/packages/heatmaps/). The first two heatmaps confirm that CNEs are enriched in AT inside but exhibit a marked depletion of AT at their borders, consistent with what is known about their biology in vertebrates [37]. Furthermore, the TAATTA motif is enriched in insect CNEs. The motif seems to be extended further by flanking A/T nucleotides. When replacing A/T (W) with G/C (S), the heatmap pattern disappears. We ask whether this is a general property of CNEs in Metazoa and, using *CNEr*, proceed to the identification of CNEs that are conserved between (a) *C. elegans* and *C. briggsae* at 100% for more than 30 nucleotides (worm CNEs, see S5 Fig B), (b) *L. variegatus* and *S. purpuratus* at 100% for more than 50 nucleotides (sea urchin CNEs, see S5 Fig C). We observe that the same pattern does not hold in those cases, i.e. it appears like enrichment of CNEs in TAATTA is not a universal phenomenon but applies only to insect and vertebrate elements. Our pipeline is a powerful tool for studying the question of how and when this TAATTA-enrichments originated, as well as a multitude of related questions.

Downstream overlap analysis of CNEs reveals that they are

355

highly enriched in stem-cell regulatory elements

356

357 Finally, we demonstrate the utility of CNEr for general hypothesis generation about CNEs by

358 identifying elements highly conserved (>98% identity over 50 bp) between human and chicken

359 and performing global and pairwise overlap analyses against various genomic features using

360 the R/Bioconductor packages *LOLA* [38] and *regioneR* [39], respectively. *LOLA* allows for

361 enrichment analysis of genomic intervals using a core reference database assembled from

362 various resources of genomic data, while *regioneR* permits cross-validation of the findings

363 through pairwise overlap analyses. As evident from inspection of S2 Table and Fig 6, both

364 packages converge to the conclusion that the identified CNEs between human and chicken are

365 significantly enriched in *Sox2* and *Oct-4* (*POU5F1*) binding sites. *Sox2* and *Oct-4*, in concert

366 with Nanog, are believed to play key roles in maintaining pluripotency. This finding comes in

367 accordance with previous reports suggesting that several CNEs are enriched in classical

368 octamer motifs recognized by developmental homeobox transcription factors [40]. Nonetheless,

369 this is the first time that such an association of the most deeply conserved CNEs with key

370 pluripotency elements is reported in the literature, and we anticipate that more associations of

371 this kind will be revealed in the future by coupling *CNEr* with other R/Bioconductor packages.

372

373 **Fig 6: Pairwise overlap analysis of CNEs demonstrates association with *Sox2*, *POU5F1***

374 **and Nanog binding regions.**

375 In all three cases, permutation tests with 1000 permutations of CNEs are shown. In grey the

376 number of overlaps of the randomized regions with the test regions of interest (in this case,

377 *Sox2*, *POU5F1* and *Nanog*) are depicted. Those overlaps of the randomized regions cluster

378 around the black bar that represents the mean. In green, the number of overlaps of the actual

17

379   regions (CNEs in this case) with the test regions is shown and is proved to be much larger than

380   expected in all cases. The red line denotes the significance limit.

Availability and Future Directions

382   The *CNEr* package with self-contained UCSC Kent's utility source code is available at

383   Bioconductor release branch http://bioconductor.org/packages/CNEr/. Active development and

384   bug reports is hosted on github https://github.com/ge11232002/CNEr/. Currently the *BLAT* is the

385   only supported aligner for identifying repeats. Other high performance short read aligners that

386   run in parallel, such as Bowtie1/2 and BWA, are desired for large set of CNEs. Furthermore,

387   integration of GRB identification approach and GRB target gene prediction is planned for future

388   development.

Declarations

Competing interests

391   The authors declare that they have no competing interests.

Funding

Authors' contributions

GT and BL conceived the project. GT implemented the software, analyzed the data. DP tested the software, analyzed data. GT, DP, BL wrote the manuscript.

Acknowledgements

We are grateful to the Bioconductor community for trying out the *CNEr* package and providing useful input.

Supporting information

S1 Text. Glossina and sea urchin data.

S2 Text. Working with Paired Genomic Ranges

S1 Fig. The heatmap shows the percentage of matched bases in the Axt alignments.

This can be useful for examining the quality of Axt alignments, especially from the whole genome pairwise alignment pipeline in *CNEr* package. The left panel has higher rates of matches than right panel since the divergence of human and mouse is much smaller than that between human and zebrafish.

411 ## S2 Fig. The syntenic plot of alignment blocks between chr1, chr2
412 ## of human and chr1, chr2 of mouse.

413 This plot is mostly used for tuning the parameters during whole genome pairwise alignment to

414 get better alignments. It can also show ancient duplications for the alignment of a sequence

415 against itself.

416 ## S3 Fig. The distribution of CNEs along the 6 biggest
417 ## chromosomes in human genome.

418 Each CNE is plotted as a dot with the position in chromosome as x-axis. A sharp increase in y-

419 axis represents a CNE cluster.

420 ## S4 Fig. The species tree of Drosophila, Glossina and mosquitos.

421 The phylogenetic tree is constructed based on the data on last common ancestors from

422 TimeTree (Hedges et al., 2006). The genome of the malaria mosquito *A. gambiae* is highly

423 divergent from *Drosophila* family and unsuitable for comparative genomics study, while *G.*

424 *morsitans* is much closer.

425 ## S5 Fig . Sequence heatmaps of CNEs in different lineages.

426 (A) *D. melanogaster* and *D. virilis* (B) *C. elegans* and *C. briggsae* (C) *L. variegatus* and *S.*

427 *purpuratus.* The CNEs are ranked by decreasing CNE width.

428 # S1 Table. A list of the most prominent CNE clusters detected

429 # between *Drosophila* and *Glossina*.

430 # S2 Table. Overlap analysis of CNEs.

431 Global overlap analysis of CNEs against multiple genomic features using LOLA reveals that

432 they overlap with Nanog, Sox2 and POU5F1 binding regions. This table is the output of

433 runLOLA algorithm sorted by FDR. All top hits include important factors associated with

434 pluripotency. userSet: all CNEs conserved between human and chicken over 98% identity for

435 more than 50 bp. collection: all elements in CODEX database. universe: all active DNase I

436 hypersensitive sites.

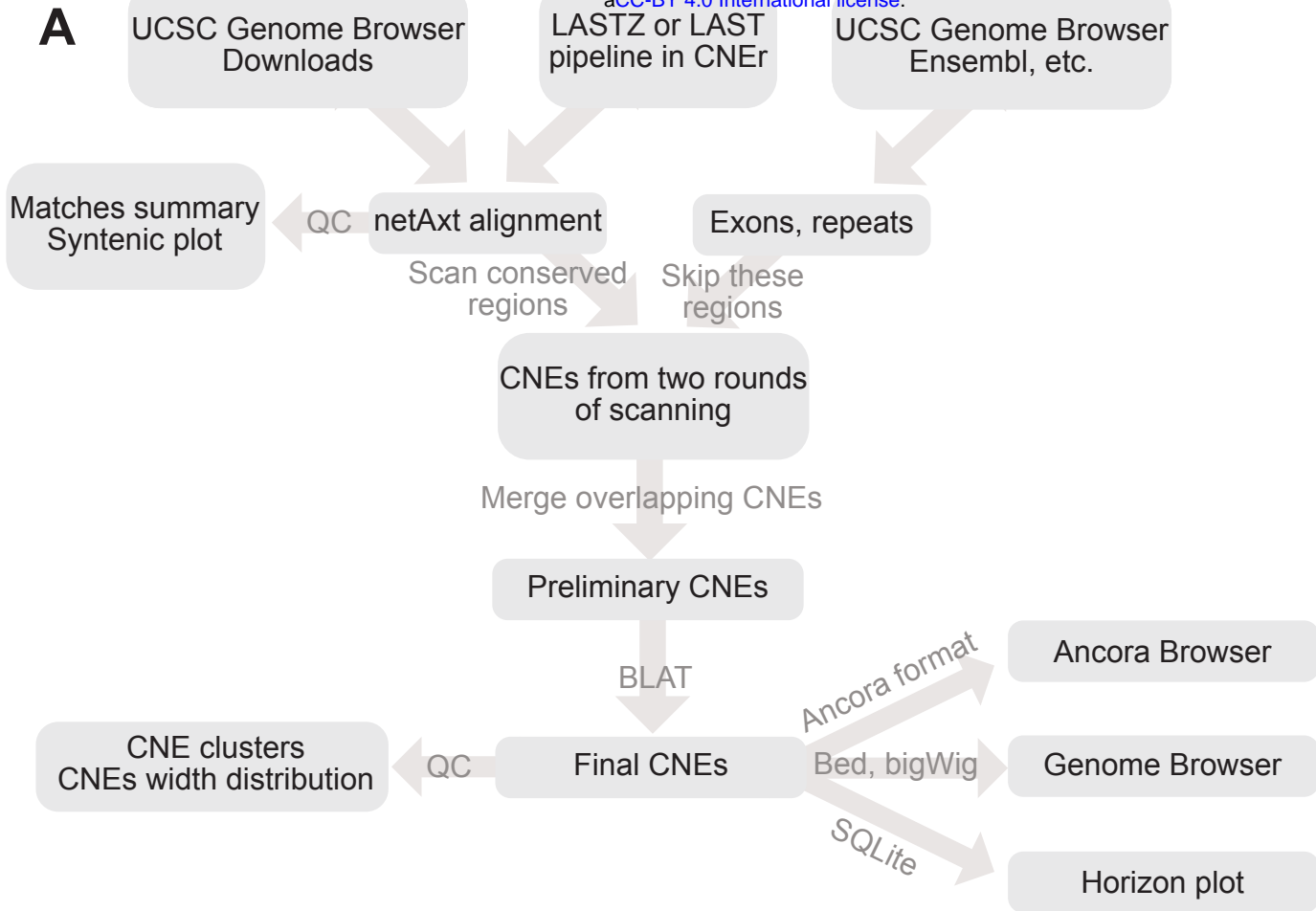437 # S1 Algorithm. The algorithm of scanning Axt alignment.

438

439 **References:**

440 1. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, et al. Highly conserved
441 non-coding sequences are associated with vertebrate development. PLoS Biol. 2005;3: e7.
442 2. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, et al. Ultraconserved
443 elements in the human genome. Sci. 2004;304: 1321–1325.
444 3. de la Calle-Mustienes E, Feijóo CG, Manzanares M, Tena JJ, Rodríguez-Seguel E, Letizia
445 A, et al. A functional survey of the enhancer activity of conserved non-coding sequences
446 from vertebrate Iroquois cluster gene deserts. Genome Res. 2005;15: 1061–1072.
447 4. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, et al. In vivo
448 enhancer analysis of human conserved non-coding sequences. Nature. 2006;444: 499–
449 502.
450 5. Harmston N, Baresic A, Lenhard B. The mystery of extreme non-coding conservation.
451 Philos Trans R Soc London Ser B, Biol Sci. 2013;368: 20130021.
452 6. Polychronopoulos D, King JWD, Nash AJ, Tan G, Lenhard B. Conserved non-coding
453 elements: developmental gene regulation meets genome organization. Nucleic acids Res.
454 2017;45: 12611–12624.
455 7. Engström PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B. Genomic regulatory blocks
456 underlie extensive microsynteny conservation in insects. Genome Res. 2007;17: 1898–
457 1908.
458 8. Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, Fredman D, et al.
459 Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved

460   synteny in vertebrates. Genome Res. 2007;17: 545–555.

461   9.   Harmston N, Ing-Simmons E, Tan G, Perry M, Merkenschlager M, Lenhard B. Topologically
462   associating domains are ancient features that coincide with Metazoan clusters of extreme
463   noncoding conservation. Nat Commun. 2017;8: 441.

464   10.   Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, et al. Enhancer
465   connectome in primary human cells identifies target genes of disease-associated DNA
466   elements. Nat Genet. 2017;49: 1602–1612.

467   11.   Montalbano A, Canver MC, Sanjana NE. High-Throughput Approaches to Pinpoint Function
468   within the Noncoding Genome. Mol cell. 2017;68: 44–59.

469   12.   Wright JB, Sanjana NE. CRISPR Screens to Discover Functional Noncoding Elements.
470   Trends Genet TIG. 2016;32: 526–529.

471   13.   Diao Y, Li B, Meng Z, Jung I, Lee AY, Dixon J, et al. A new class of temporarily phenotypic
472   enhancers identified by CRISPR/Cas9-mediated genetic screening. Genome Res. 2016;26:
473   397–405.

474   14.   Sanjana NE, Wright J, Zheng K, Shalem O, Fontanillas P, Joung J, et al. High-resolution
475   interrogation of functional elements in the noncoding genome. Sci. 2016;353: 1545–1549.

476   15.   Royo JL, Bessa J, Hidalgo C, Fernández-Miñán A, Tena JJ, Roncero Y, et al. Identification
477   and analysis of conserved cis-regulatory regions of the MEIS1 gene. PloS one. 2012;7:
478   e33617.

479   16.   Anderson E, Devenney PS, Hill RE, Lettice LA. Mapping the Shh long-range regulatory
480   domain. Dev. 2014;141: 3934–3943.

481   17.   Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with
482   space/time models. Briefings Bioinforma. 2011;12: 41–51.

483   18.   Ayad LAK, Pissis SP, Polychronopoulos D. CNEFinder: finding conserved non-coding
484   elements in genomes. Bioinformatics (Oxford, England). 2018. pp. i743–i747.

485   19.   Sandelin A, Bailey P, Bruce S, Engström PG, Klos JM, Wasserman WW, et al. Arrays of
486   ultraconserved non-coding regions span the loci of key developmental genes in vertebrate
487   genomes. BMC Genomic-. 2004;5: 99.

488   20.   Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human
489   genome browser at UCSC. Genome Res. 2002;12: 996–1006.

490   21.   Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016.
491   Nucleic acids Res. 2016;44: D710–D716.

492   22.   Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use
493   annotation pipeline designed for emerging model organism genomes. Genome Res.
494   2008;18: 188–196.

495   23.   Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, et al. Genome
496   duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-
497   karyotype. Nature. 2004;431: 946–957.

498   24.   Kolder ICRM, van der Plas-Duivesteijn SJ, Tan G, Wiegertjes GF, Forlenza M, Guler AT, et
499   al. A full-body transcriptome and proteome resource for the European common carp. BMC
500   Genomic-. 2016;17: 701.

501   25.   Hahne F, Ivanek R. Visualizing Genomic Data Using Gviz and Bioconductor. Methods Mol
502   Biol. 2016;1418: 335–351.

503   26.   Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12: 656–664.

504   27.   Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, et al. Human-mouse
505   alignments with BLASTZ. Genome Res. 2003;13: 103–107.

506   28.   Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence
507   comparison. Genome Res. 2011;21: 487–493.

508   29.   Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication,

509   deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci United
510   States Am. 2003;100: 11484–11489.
511 30. International Glossina Genome Initiative. Genome sequence of the tsetse fly (Glossina
512   morsitans): vector of African trypanosomiasis. Sci. 2014;344: 380–386.
513 31. Engström PG, Fredman D, Lenhard B. Ancora: a web resource for exploring highly
514   conserved noncoding elements and their association with developmental regulatory genes.
515   Genome Biol. 2008;9: R34.
516 32. Reim I, Mohler JP, Frasch M. Tbx20-related genes, mid and H15, are required for tinman
517   expression, proper patterning, and normal differentiation of cardioblasts in Drosophila.
518   Mech Dev. 2005;122: 1056–1069.
519 33. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological
520   themes among gene clusters. Omics: J Integr Biol. 2012;16: 284–287.
521 34. Nepveu A. Role of the multifunctional CDP/Cut/Cux homeodomain transcription factor in
522   regulating differentiation, cell growth and development. Gene. 2001;270: 1–15.
523 35. Cameron RA, Samanta M, Yuan A, He D, Davidson E. SpBase: the sea urchin genome
524   database and web site. Nucleic acids Res. 2009;37: D750–D754.
525 36. Chiang CWK, Derti A, Schwartz D, Chou MF, Hirschhorn JN, Wu C-T. Ultraconserved
526   elements: analyses of dosage sensitivity, motifs and boundaries. Genetics. 2008;180:
527   2277–2293.
528 37. Walter K, Abnizova I, Elgar G, Gilks WR. Striking nucleotide frequency pattern at the
529   borders of highly conserved vertebrate non-coding sequences. Trends Genet TIG. 2005;21:
530   436–440.
531 38. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory
532   elements in R and Bioconductor. Bioinforma. 2016;32: 587–589.
533 39. Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. regioneR: an
534   R/Bioconductor package for the association analysis of genomic regions based on
535   permutation tests. Bioinforma. 2016;32: 289–291.
536 40. Chiang CWK, Liu C-T, Lettre G, Lange LA, Jorgensen NW, Keating BJ, et al.
537   Ultraconserved elements in the human genome: association and transmission analyses of
538   highly constrained single-nucleotide polymorphisms. Genetics. 2012;192: 253–266.

Figure 1

A

UCSC Genome Browser Downloads

LASTZ or LAST pipeline in CNEr

UCSC Genome Browser Ensembl, etc.

Matches summary Syntenic plot

QC netAxt alignment

Exons, repeats

Scan conserved regions

Skip these regions

CNEs from two rounds of scanning

Merge overlapping CNEs

Preliminary CNEs

BLAT

Ancora format → Ancora Browser

CNE clusters CNEs width distribution

QC Final CNEs

Bed, bigWig → Genome Browser

SQLite → Horizon plot

B Axt alignment

```
TGTTCCTATAGCTAAAAGCATAAC-GGCTTTGATTACATGCTAAATAGAAT
TGT-ACTTTAGCTAAAA-CATAATAGGCTTTGATTACATGCTAAATAGAAT
***   ** ******** *****  **************************
```
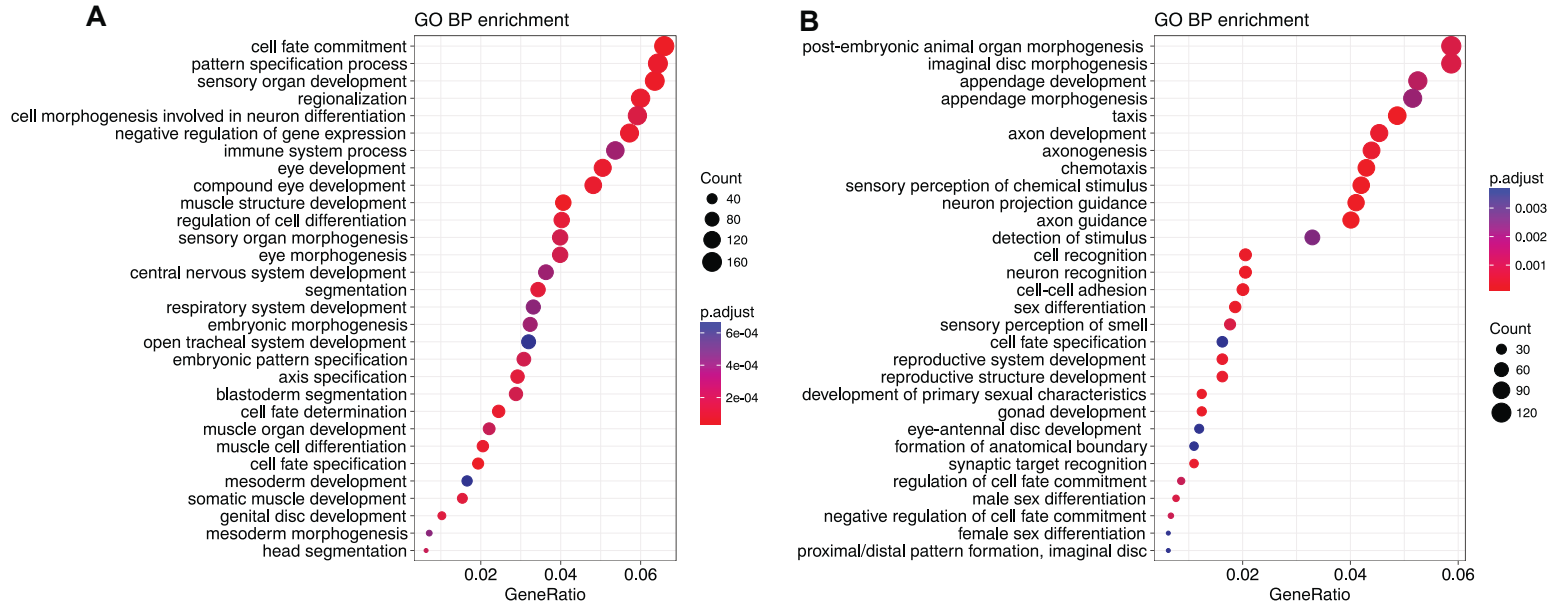
Below the threshold   Exons   CNE   Repeats

Figure 2

Figure 3

Figure 4

Figure 5

Figure 6