D96–D100 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 2 November 2015*
doi: 10.1093/nar/gkv1163

CEGA—a catalog of conserved elements from genomic alignments

Aline Dousse, Thomas Junier and Evgeny M. Zdobnov*

Department of Genetic Medicine and Development, University of Geneva Medical School, Swiss Institute of Bioinformatics, rue Michel-Servet 1, 1211 Geneva, Switzerland

Received August 14, 2015; Revised September 25, 2015; Accepted October 20, 2015

ABSTRACT

By identifying genomic sequence regions conserved among several species, comparative genomics offers opportunities to discover putatively functional elements without any prior knowledge of what these functions might be. Comparative analyses across mammals estimated 4–5% of the human genome to be functionally constrained, a much larger fraction than the 1–2% occupied by annotated protein-coding or RNA genes. Such functionally constrained yet unannotated regions have been referred to as conserved non-coding sequences (CNCs) or ultra-conserved elements (UCEs), which remain largely uncharacterized but probably form a highly heterogeneous group of elements including enhancers, promoters, motifs, and others. To facilitate the study of such CNCs/UCEs, we present our resource of Conserved Elements from Genomic Alignments (CEGA), accessible from http://cega.ezlab.org. Harnessing the power of multiple species comparisons to detect genomic elements under purifying selection, CEGA provides a comprehensive set of CNCs identified at different radiations along the vertebrate lineage. Evolutionary constraint is identified using threshold-free phylogenetic modeling of unbiased and sensitive global alignments of genomic synteny blocks identified using protein orthology. We identified CNCs independently for five vertebrate clades, each referring to a different last common ancestor and therefore to an overlapping but varying set of CNCs with 24 488 in vertebrates, 241 575 in amniotes, 709 743 in Eutheria, 642 701 in Boreoeutheria and 612 364 in Euarchontoglires, spanning from 6 Mbp in vertebrates to 119 Mbp in Euarchontoglires. The dynamic CEGA web interface displays alignments, genomic locations, as well as biologically relevant data to help prioritize and select CNCs of interest for further functional investigations.

INTRODUCTION

Genome sequencing provides access to the complete repertoire of inherited functional elements, from encoded genes to regulatory sequences, but recognizing these elements and understanding their biological activities remains challenging. Comparative genomics offers an approach to help recognize such elements, by identifying sequences that remain conserved across multiple species over millions of years of evolution (1). Their intolerance to mutations, making them appear as conserved, implies functional constraints on such sequences, regardless of our knowledge of their functions. Applying such methods to the increasing number of sequenced genomes has helped to identify core genes conserved across many species and has additionally revealed a repertoire of genomic elements at least as large as that of protein-coding genes that does not encode proteins or RNA genes (2–5). These elements were termed Conserved Non-Coding sequences (CNCs) (6,7), and the investigation of their functional roles is still ongoing. Some of the most highly-conserved elements in vertebrates, termed Ultra-Conserved Elements (UCEs) have been tested *in vivo*, but only about half of these showed any capacity for specific *cis*-regulatory activities (8).

Beyond the lack of systematic experimental investigations of such CNCs or UCEs, the variable technical definitions used to classify such elements have hampered progress in this field of research. For example, the working definition of CNCs from pioneering studies (9) selected an arbitrary threshold of a minimum sequence identity over a minimum alignment length in pairwise sequence comparisons, which is still a frequently used definition. However, with no systematic approach to select threshold parameters, the results from employing such a definition are clearly impacted by the evolutionary distance between the pair of species being compared. Various strategies have been developed to fine-tune these definitions in order to generate genome-scale resources of computationally-identified CNCs and to help prioritize candidates in turn satisfying the growing interest in developing functional screens of these elements. Most of the existing resources employ pairwise DNA alignments as a starting point to define CNCs, e.g. human and

*To whom correspondence should be addressed. Tel: +41 22 379 59 73; Fax: +41 22 379 57 06; Email: evgeny.zdobnov@unige.ch mouse (VISTA enhancer browser (8)), human and chicken (UCNEbase (10)), human and zebrafish (cneViewer, (11)), or human and fugu (CONDOR (12)). However, pairwise alignments lack comparative power, and ignore the additional evolutionary information to be gleaned from including any of the dozens of vertebrate genomes already sequenced. Extending such approaches by searching 'seed' CNCs identified from pairwise comparisons to other species does not completely resolve this issue. If a conserved element is not present (or sequenced) in the species chosen for the initial pairwise comparison, it will not be part of the final set of CNCs. Thus, pairwise approaches are inherently biased and not comprehensive. Similarly, the extension of the pairwise approach to several species, by the choice of a reference organism and subsequent alignments to it, is also biased and not very sensitive to distantly-related species. In addition, definitions of conservation vary considerably, e.g. 100% identity over ≥200 bp for the VISTA enhancer browser, ≥95% identity over ≥200 bp for UCNEbase, user-defined conservation, length, or distance cutoffs for cneViewer, 70–100% identity over ≥30 bp or ≥50 bp for ANCORA (13) (depending on the species pair being considered), ≥70% identity over ≥100 bp (human–mouse) and ≥65% identity over ≥50 bp (mammal–fugu) for TFCONES, ≥65% identity over ≥40 bp for CONDOR and 100% identity over ≥200 bp in human, mouse and rat for UCbase2.0 (14). Some of these resources additionally provide access to the results of regulatory screening of CNCs. Notably, the Vista enhancer browser indexes the results of gene enhancer activity in transgenic mice for 2192 elements, and around a hundred of the 7000 CNCs in CONDOR were tested *in-vivo* for enhancer activity in zebrafish embryos.

To advance the field of research on the identification and characterization of CNCs, we devised a computational pipeline to yield a comprehensive and unbiased set of conserved elements at different radiations along the vertebrate lineage. We overcome the limitations of pairwise approaches by harnessing the power of multiple species comparisons to detect genomic elements under purifying selection using at least five species (15). We gain sensitivity for identifying CNCs by employing global sequence alignments without a reference organism (16) for each of the collinear genomic blocks defined by protein orthology. We rely on phastCons (17) phylogenetic modeling to objectively define conserved elements. The resulting catalog of Conserved Elements from Genomic Alignments (CEGA) provides access to these sets of conserved elements from http://cega.ezlab.org, 24 488 CNCs in the vertebrate clade to 612 364 CNCs in the euarchontoglires (Supraprimates). The CEGA web interface allows browsing of all conserved elements annotated as coding, intergenic or intronic with complementary features selected from the The Encyclopedia of DNA Elements (ENCODE) data (18) such as chromatin state annotations (19) which provide clues to their possible biological functions.

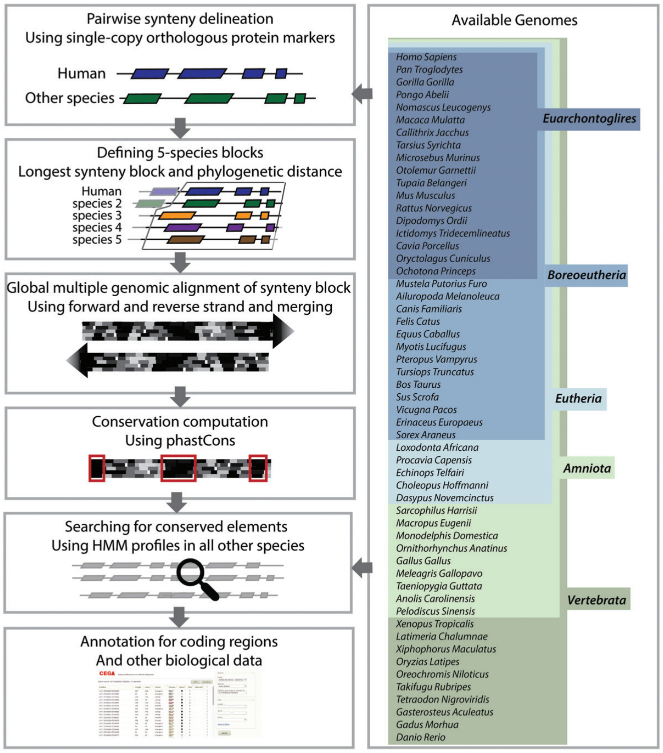


Figure 1. Workflow of CEGA identification of conserved elements.

IDENTIFICATION OF CONSERVED GENOMIC ELEMENTS

The CEGA resource presents sets of conserved genomic elements computed independently from a total of 55 vertebrate species. Conservation of a genomic element in a set of species implicitly refers to its presence in the last common ancestor (LCA) of these species. We therefore independently considered five different vertebrate clades: vertebrates, amniotes, Eutheria, Boreoeutheria and Euarchontoglires, each referring to a different LCA and thus to an overlapping but varying set of CNCs. With a total of 1 398 498 CNCs, CEGA offers a comprehensive catalog of conserved elements at each level of the vertebrate phylogeny (Table 1), a third of which are intergenic and the remaining two thirds are intronic. The steps that comprise CEGAs computational pipeline to identify CNCs are explained below and in further details in the Supplementary Material as well as schematically outlined in Figure 1.

Synteny block delineation

The CEGA pipeline starts with the identification of collinear genomic blocks, also termed synteny blocks, to then be able to perform reliable alignments of the sequences from each block. Delineation of these blocks is based on single-copy orthologous protein markers from OrthoDB (release7) (20). Protein-based markers provide the advantages of having a slower rate of sequence evolution, additional informational content at the amino acid level, and longer sequences than DNA markers used by other approaches to identify orthologous relations and synteny blocks, e.g. in Enredo (21). After looking for synteny blocks between pairs of species, CEGA defines blocks with sets of five species, sufficient to harness the comparative power (15). To maximize the coverage of synteny blocks across the human genome and to take advantage of the growing number of available vertebrate genomes, each block may be de-

Table 1. CEGA data content

	Vertebrata	Amniota	Eutheria	Boreoeutheria	Euarchontoglires
Input species	55	45	36	31	18
Included species	43	42	36	31	18
Synteny blocks	1649	1880	1713	1319	1326
Synteny block length[a] [Mb]	607	1479	1677	1763	1830
Conserved elements	66 280	361 876	869 050	801 032	742 702
CNCs	24 488	241 575	709 743	642 701	612 364
Median CNCs length [bp]	190	147	108	116	128
Total CNCs length [Mb]	6	52	122	116	119

[a] Total synteny block length across the human genome.

fined by different sets of species. In practice, we developed a scoring system based on the phylogenetic distance between each pair of species, the length of the block in terms of number of orthologous protein markers, and the genome sequence quality in terms of gaps in the assemblies to automatically select the best combination of species. Species selection is constrained to contain human and at least one organism from the root level (i.e. most distant from human) of the investigated clade in order to fully span the phylogeny. The common markers identified in all five pairwise blocks across are extracted and the corresponding genomic sequence is further extended by 15 Kb flanks in each genome to include additional intergenic sequences. Employing this strategy resulted in large fractions (from 608 Mb to 1’830 Mb) of the human genome being delineated in to synteny blocks (Table 1).

Multiple sequence alignments

Although we focused on human, requiring all synteny blocks to contain human genomic sequence, CEGA aims to provide unbiased alignments, without using a reference organism. We also target conservation across large evolutionary distances (human – fish) requiring sensitive alignments. Since the protein-orthology-based identification of synteny blocks implies the orthology of the corresponding genomic regions, we opted to use global alignment approaches that attempt to align sequences along the whole lengths of the genomic sequences. After extensive benchmarking of many available alignment methods, we selected MLAGAN (16) that provides global alignments with a local anchoring strategy without requiring the selection of a reference species. To overcome the ‘Heads or Tails’ bias problem (22), i.e. obtaining different results when aligning the exact same sequences in the forward and reverse orientations, we aligned each of the five sequences of the synteny blocks both on the forward and the reverse strands and reconciled the alignments by merging them with MergeAlign (23).

Identification of conservation

To avoid selecting arbitrary identity and lengths thresholds, the classification of CEGA sets of conserved element employs phylogenetic modeling with phastCons (17) to define evolutionarily constrained elements. The conservation metrics reported are log-odds scores of the probability of the element following a conserved model rather than an unconserved model (parameters and models are described in detail in the supplementary material). Finally, elements with ‘N’ sequence stretches or having less than 20 nucleotides aligned were filtered out of the conserved set.

Expanding to other species

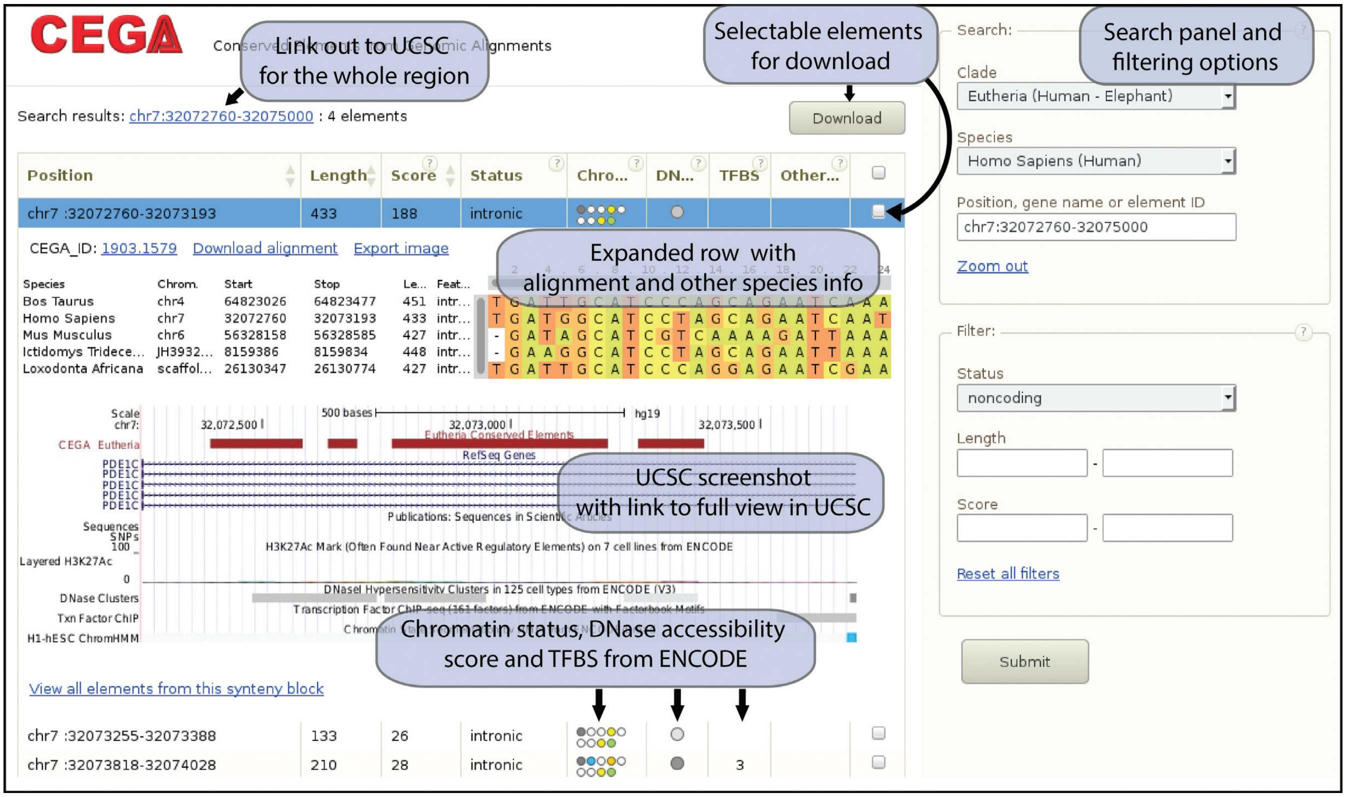
Since conserved elements are initially identified from a subset of only five species, we used a hidden Markov model (HMM) profile built with *nhmmer* from HMMER 3.1 (24) using each individual element to search whole genome assemblies of all other species. These searches were carried out using the set of vertebrate elements to search all vertebrate genomes. A similar strategy was used for the amniote clade, but limiting the searches to those synteny blocks with previously identified orthologous regions in other species. The highest scoring significant match (e-value <0.05) was selected and the alignment of the element recomputed using muscle (25). Currently, CEGA only presents HMM-expanded elements for Vertebrate and Amniotes. In further development the same procedure can be applied to the other clades.

Functional annotation

Based on Ensembl gene annotations (26) for all species, elements were annotated as either protein-coding, RNA-coding (micro-RNA and long non-coding RNA), intronic or intergenic. These classifications are complemented with selected annotation from the ENCODE project (18) were incorporated such as the chromatin state (19), the number of transcription factors that bind to the genomic region (27), and DNase accessibility values (28). These represent informative annotations that offer additional evidence to help select elements for future investigations of their biological function investigations, e.g. *cis*-regulatory activities. In addition, overlapping ultra-conserved elements defined by alternative approaches available from other databases are listed for easy cross-referencing.

CEGA DATABASE CONTENT

CEGA database is structured into five main data tables, for each of the investigated vertebrate clade: Vertebrates, Amniotes, Eutheria, Boreoeutheria and Euarchontoglires. In each table, conserved elements are organized into synteny blocks, and then per element. Each element has an ID and information about its location in each species, as well as its


CEGA
Conserved Elements from Genomic Alignments
Link out to UCSC for the whole region
Selectable elements for download
Search panel and filtering options
Expanded row with alignment and other species info
UCSC screenshot with link to full view in UCSC
Chromatin status, DNase accessibility score and TFBS from ENCODE
Download
Zoom out
Reset all filters
Submit
Scale
chr7:
32,072,500
500 bases
hg19
32,073,000
32,073,500
Eutheria Conserved Elements
CEGA Eutheria
RefSeq Genes
PDE1C
PDE1C
PDE1C
PDE1C
PDE1C
Publications: Sequences in Scientific Articles
Sequences
SNPs
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNase Clusters
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
Chromatin State Segmentation by HMM from ENCODE/Broad
H1-hESC ChromHMM
2 4 6 8 10 12 14 16 18 20 22 24

Search results: chr7:32072760-32075000 : 4 elements

Position	Length	Score	Status	Chro...	DN...	TFBS	Other...	□
chr7 :32072760-32073193	433	188	intronic	●○○●○ ○○●●	○			□

CEGA_ID: 1903.1579 Download alignment Export image

Species	Chrom.	Start	Stop	Le...	Feat...	
Bos Taurus	chr4	64823026	64823477	451	intr...	T G A T T G C A T C C C A G C A G A A T C A A A
Homo Sapiens	chr7	32072760	32073193	433	intr...	T G A T G G C A T C C T A G C A G A A T C A A T
Mus Musculus	chr6	56328158	56328585	427	intr...	- G A T A G C A T C G T C A A A A G A T T A A A
Ictidomys Tridece...	JH3932...	8159386	8159834	448	intr...	- G A A G G C A T C C T A G C A G A A T T A A A
Loxodonta Africana	scaffol...	26130347	26130774	427	intr...	T G A T T G C A T C C C A G G A G A A T C G A A

View all elements from this synteny block

chr7 :32073255-32073388	133	26	intronic	●○○●○ ○○●●	○			□
chr7 :32073818-32074028	210	28	intronic	●●○●○ ○○●●	●	3		□

Search:

Clade: Eutheria (Human - Elephant)

Species: Homo Sapiens (Human)

Position, gene name or element ID: chr7:32072760-32075000

Filter:

Status: noncoding

Length: ___ - ___

Score: ___ - ___

Figure 2. CEGA user interface.

sequence and the corresponding annotations. The synteny blocks cover from 20% to 63% of the human genome, with CNCs ranging from 1% to 6.5% of these blocks, depending on the level of the vertebrate phylogeny.

CEGA web interface

The database is accessible through a dynamic web interface browsable by selecting a region of interest on a human chromosome view or by submitting a genomic location or a gene of interest. In the latter case, the gene genomic position is retrieved from annotations (26) and further expanded with 1 Mbp flanks. The main CEGA display is a table showing the previously described information about each conserved element overlapping the submitted locus. Each row is expandable by just one click to view the sequence alignment from the element, its details in the other species and a screenshot of the genomic location of the element from the UCSC genome browser (29). Access to UCSC browser displaying CEGA tracks can be made directly from the whole region table or from the element view. This functionality is intended as an entry point for the analysis of further biologically-relevant annotations.

As shown on the example of CEGA interface on Figure 2, three columns of the table represent selected biological data from ENCODE that can help to make a selection of relevant elements. The DNAse column shows the DNase sensitivity of the locus with a grayscale, from white for no data or 0 scoring to black for the highest score. This score is based on the combination of the DNAse sensitivity in 125 cell-types (28). Regulatory regions are usually DNase sensitive. The regulatory potential of the element is further detailed by the chromatin state column. Nine circles, one for each of the investigated cell lines, are colored according to the type of activity the integration of chromatin marks data (19) suggests for the genomic region; warm colors represent promoter and enhancer regions whereas cold ones suggest repressed or repetitive region and heterochromatin. The TFBS column simply shows the number of transcription factors with a ChipSeq peak overlapping the conserved element in any of the tested cell lines. This number allows for the selection of highly interacting element over others. A last column is dedicated to the overlap of each CEGA element with ultra-conserved elements identified by other methods. These databases provide other information: gene regulatory blocks and potential gene regulated in UCNE and experimental annotations in CONDOR and Vista enhancer browser.

A checkbox allows the user to select its elements of interest and get bed or Fasta files for them. Bed files can be used as UCSC tracks, looking for overlaps with specific markers and Fasta is provided to look for similar elements or to explore their evolutionary history. The complete set of CEGA data is also available for bulk download.

CONCLUSIONS AND PERSPECTIVES

CEGA aims to provide an easy access to unbiased and comprehensive sets of CNCs at distinct levels of the vertebrate lineage. The sets were computed based on a strategy to be as comprehensive and sensitive as possible, while keeping scalability in mind. The strategy of using five species per block can cope with the rapidly increasing number of sequenced genomes while harnessing the comparative power. In the future more species can be included without becoming a computational hurdle. This method has however a drawback of not presenting a constant collection of species, not all conserved elements were computed in all species. CEGA provides a convenient access using dynamic webpages to all elements within a genomic interval or close to a particular gene. Quick visualization of relevant biological data in relation to the conserved elements is also provided and can help prioritize the in-depth investigation of a sub-group of elements. Therefore elements can be selected and downloaded in various formats: as bed-file for visualization and for finding overlaps with other features, as multiple alignments in Fasta format for phylogenetic studies or single sequence Fasta for further studies and comparisons.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors thank all members of the Zdobnov group for useful discussions and suggestions.

FUNDING

Swiss National Science Foundation [PDFMP3_127347]; Swiss Institute of Bioinformatics (SIB) and Foundation Cavaglieri (in part).

Conflict of interest statement. None declared.

REFERENCES

1. Alföldi,J. and Lindblad-Toh,K. (2013) Comparative genomics as a tool to understand evolution and disease. *Genome Res.*, **23**, 1063–1068.
2. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
3. Dermitzakis,E.T., Reymond,A., Lyle,R., Scamuffa,N., Ucla,C., Deutsch,S., Stevenson,B.J., Flegel,V., Bucher,P., Jongeneel,C.V. *et al.* (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature*, **420**, 578–582.
4. Dubchak,I., Brudno,M., Loots,G.G., Pachter,L., Mayor,C., Rubin,E.M. and Frazer,K.A. (2000) Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.*, **10**, 1304–1306.
5. Dermitzakis,E.T., Reymond,A. and Antonarakis,S.E. (2005) Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat. Rev. Genet.*, **6**, 151–157.
6. Couronne,O., Poliakov,A., Bray,N., Ishkhanov,T., Ryaboy,D., Rubin,E., Pachter,L. and Dubchak,I. (2003) Strategies and tools for whole-genome alignments. *Genome Res.*, **13**, 73–80.
7. Drake,J.A., Bird,C., Nemesh,J., Thomas,D.J., Newton-Cheh,C., Reymond,A., Excoffier,L., Attar,H., Antonarakis,S.E., Dermitzakis,E.T. *et al.* (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat. Genet.*, **38**, 223–227.
8. Visel,A., Minovitsky,S., Dubchak,I. and Pennacchio,L.A. (2007) VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, 88–92.
9. Duret,L., Dorkeld,F. and Gautier,C. (1993) Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. *Nucleic Acids Res.*, **21**, 2315–2322.
10. Dimitrieva,S. and Bucher,P. (2013) UCNEbase–a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res.*, **41**, D101–D109.
11. Persampieri,J., Ritter,D.I., Lees,D., Lehoczky,J., Li,Q., Guo,S. and Chuang,J.H. (2008) cneViewer: a database of conserved non-coding elements for studies of tissue-specific gene regulation. *Bioinformatics*, **24**, 2418–2419.
12. Woolfe,A., Goode,D.K., Cooke,J., Callaway,H., Smith,S., Snell,P., McEwen,G.K. and Elgar,G. (2007) CONDOR: a database resource of developmentally associated conserved non-coding elements. *BMC Dev. Biol.*, **7**, 100.
13. Engström,P.G., Fredman,D. and Lenhard,B. (2008) Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. *Genome Biol.*, **9**, R34.
14. Lomonaco,V., Martoglia,R., Mandreoli,F., Anderlucci,L., Emmett,W., Bicciato,S. and Taccioli,C. (2014) UCbase 2.0: ultraconserved sequences database (2014 update). *Database (Oxford).*, 1–8.
15. Eddy,S.R. (2005) A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.*, **3**.
16. Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., Green,E.D., Sidow,A. and Batzoglou,S. (2003) LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.*, **13**, 721–731.
17. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
18. Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
19. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
20. Waterhouse,R.M., Tegenfeldt,F., Li,J., Zdobnov,E.M. and Kriventseva,E. V (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.*, **41**, D358–D365.
21. Paten,B., Herrero,J., Beal,K., Fitzgerald,S. and Birney,E. (2008) Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.*, **18**, 1814–1828.
22. Landan,G. and Graur,D. (2007) Heads or tails: a simple reliability check for multiple sequence alignments. *Mol. Biol. Evol.*, **24**, 1380–1383.
23. Collingridge,P.W. and Kelly,S. (2012) MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. *BMC Bioinformatics*, **13**, 117.
24. Wheeler,T.J. and Eddy,S.R. (2013) nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, **29**, 2487–2489.
25. Edgar,R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
26. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, 749–755.
27. Gerstein,M.B., Kundaje,A., Hariharan,M., Landt,S.G., Yan,K.-K., Cheng,C., Mu,X.J., Khurana,E., Rozowsky,J., Alexander,R. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91–100.
28. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
29. Rosenbloom,K.R., Armstrong,J., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Dreszer,T.R., Fujita,P., Guruvadoo,L., Haeussler,M. *et al.* (2014) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.