Ryu et al. BMC Evolutionary Biology 2012, 12:236
http://www.biomedcentral.com/1471-2148/12/236

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

The evolution of ultraconserved elements with different phylogenetic origins

Taewoo Ryu[1,2*], Loqmane Seridi[1,2] and Timothy Ravasi[1,2*]

Abstract

Background: Ultraconserved elements of DNA have been identified in vertebrate and invertebrate genomes. These elements have been found to have diverse functions, including enhancer activities in developmental processes. The evolutionary origins and functional roles of these elements in cellular systems, however, have not yet been determined.

Results: Here, we identified a wide range of ultraconserved elements common to distant species, from primitive aquatic organisms to terrestrial species with complicated body systems, including some novel elements conserved in fruit fly and human. In addition to a well-known association with developmental genes, these DNA elements have a strong association with genes implicated in essential cell functions, such as epigenetic regulation, apoptosis, detoxification, innate immunity, and sensory reception. Interestingly, we observed that ultraconserved elements clustered by sequence similarity. Furthermore, species composition and flanking genes of clusters showed lineage-specific patterns. Ultraconserved elements are highly enriched with binding sites to developmental transcription factors regardless of how they cluster.

Conclusion: We identified large numbers of ultraconserved elements across distant species. Specific classes of these conserved elements seem to have been generated before the divergence of taxa and fixed during the process of evolution. Our findings indicate that these ultraconserved elements are not the exclusive property of higher modern eukaryotes, but rather transmitted from their metazoan ancestors.

Keywords: Ultraconserved elements, Developmental enhancers, Transcriptional regulatory networks, Genome evolution, Marine biology

Background

Large numbers of DNA elements (≥200 bp) exhibiting 100% similarity have been found to be conserved across several mammalian species [1,2]. Shorter ultraconserved elements (UCEs) longer than 50 bp and 100 bp have also been identified in several insect species and plants, respectively [3,4].

Since the discovery of UCEs, a lot of effort has been expended on elucidating their functions and to determine the reasons for their extreme conservation. UCEs are often located near genes implicated in transcription and developmental processes, splicing, and ion flow control across membranes [1,2,5-7]. In vivo analysis of the embryos of transgenic mice uncovered the transcriptional enhancer activities of UCEs targeting developmental genes and transcription factors (TFs) [8,9]. Depletion of UCEs among segmental duplications and copy number variations were also reported [10]. Single nucleotide polymorphisms (SNPs) in UCEs have been linked to cancer risk, impaired TF binding, and homeobox gene regulation in the central nervous system [11,12]. Nevertheless, homozygote embryo knockout experiments in mice revealed that deletion of ultraconserved elements can yield viable mice, suggesting the dispensability or functional redundancy of UCEs [13].

The origin and evolution of UCEs have also been also investigated. There is evidence that some UCEs originated from retroposons and stabilized in genomes after acquiring a function that benefitted the host [14]. Stephen et al. studied the evolution of UCEs in several

* Correspondence: taewoo.ryu@kaust.edu.sa; timothy.ravasi@kaust.edu.sa
[1]Integrative Systems Biology Lab., Division of Biological and Environmental Sciences & Engineering, Division of Applied Mathematics and Computer Sciences, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia
[2]Department of Medicine, Division of Medical Genetics, University of California, San Diego, 9500 Gilman Drive La Jolla, CA 92093-0688, USA

vertebrate genomes and found that they were generated and expanded on a large scale during tetrapod evolution [15]. Other studies of the human genome showed that UCEs experienced strong purifying selection and were not mutational cold spots [16-18].

In this study, we investigated if evidence of the conservation of DNA elements could be found in primitive species, such as sponge and hydra, and if these conserved elements have similar functions as those previously reported for higher eukaryotes. We identified many UCEs across diverse phyla, including Porifera, Cnidaria, Arthropoda, Echinodermata, and Chordata, as well as a new type of short UCEs. By comparing distant species, we were able to identify new UCEs in human and fruit fly. Clustering the UCEs based on the sequence similarity unveiled lineage specificity and distinct functions outlined by protein domains of their flanking genes and DNA regulatory motifs. We concluded that each UCE group arose independently on a specific lineage and was "frozen" on the genome as a regulatory innovation after the divergence of specific taxa.

Results and discussion

Identification of ultraconserved elements across diverse taxa

We began our analysis by asking if there is evidence of ultraconservation in primitive species and, if so, how UCEs diverged during the process of evolution. We considered six species whose genomes were previously sequenced including demosponge (*Amphimedon queenslandica*) from the phylum Porifera, hydra (*Hydra magnipapillata*) and sea anemone (*Nematostella vectensis*) from the phylum Cnidaria, sea urchin (*Strongylocentrotus purpuratus*) from the phylum Echinodermata, fruit fly (*Drosophila melanogaster*) from the phylum Arthropoda, and human (*Homo sapiens*) from the phylum Chordata. We identified UCEs (≥50 bp) and shorter UCEs (≥30 bp) by pairwise comparison of the whole genomic sequences across six species.

Unexpectedly, the number of identified UCEs and the size of some of them (11 UCEs ≥ 200 bp) were large considering the evolutionary distance between analyzed species. This result suggested the presence of UCEs in primitive species and across distant taxa (Table 1 and Figure 1). Most of the UCEs were found in hydra and sea anemone, which belong to the same phylum, Cnidaria. However, the exact reason for the predominance of UCEs in these species cannot be addressed until more genome sequences of species around this lineage become available and current genome assemblies are improved. Interestingly, the longest UCE (796 bp) was conserved in both sea anemone and human, two species that diverged approximately 892 million years ago [19]. We found that the number of UCEs and the evolutionary distance (Table 1 and Figure 1) between species are negatively correlated, an observation that is also the case for shorter UCEs.

We noticed that a large number of conserved DNA elements that we identified overlapped in each species because the UCE-identification program, MUMmer, reported all maximal matches regardless of the overlap [20]. To minimize redundancy and facilitate downstream analysis, neighboring UCEs and short UCEs in each species were joined as non-overlapping ultraconserved regions (UCRs) (Additional file 1 and Additional file 2). The numbers of these non-overlapping UCRs (≥50 bp) were 30 for sponge, 64 for fruit fly, 673 for hydra, 56 for human, 3,807 for sea anemone, and 187 for sea urchin.

Novel ultraconserved elements in human and fruit fly

As a benchmark for our UCE discovery pipeline, we examined how many UCEs that had been previously identified we were able to recover. Previously reported UCEs in human and fruit fly were aligned to their reference genome using Bowtie [21] to determine their exact locations in the current genome build (hg19 and dm3, respectively). The majority of known UCEs (all 481 elements from the human-mouse-rat alignment [1], 23,695 out of 23,699 elements from the *D. melanogaster-Drosophila pseudoobscura* alignment, and all 126 elements from the *D. melanogaster-Anopheles gambiae* alignment [3]) were successfully aligned. We then compared these elements

Table 1 Identification of UCEs

	A. queenslandica (sponge)	*N. vectensis* (sea anemone)	*H. magnipapillata* (hydra)	*D. melanogaster* (fruit fly)	*S. purpuratus* (sea urchin)	*H. sapiens* (human)
A. queenslandica	-	2,135	669	43	108	9
N. vectensis	5,303	-	54,732	256	5,525	10
H. magnipapillata	1,300	97,669	-	125	400	0
D. melanogaster	75	5,440	478	-	188	27
S. purpuratus	537	43,707	5,498	1,129	-	19
H. sapiens	83	381	328	415	967	-

Columns and rows are sorted by the phylogeny as shown in Figure 1. Upper and lower triangles show the numbers of 100 % identical matches ≥ 50 bp and ≥ 30 bp between two species, respectively.

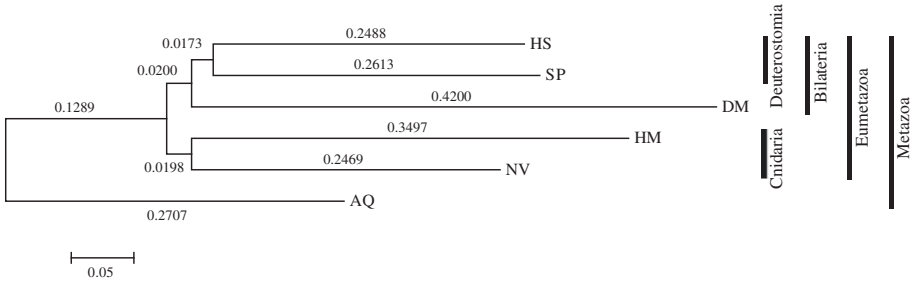

Figure 1 Evolutionary relationships between analyzed species. The JTT matrix-based method [61] is used to compute the evolutionary distances and the phylogenetic tree is constructed using the Neighbor-Joining method [62]. Bootstrapping values from 500 replicates are shown and selected taxon information is depicted on the right. Species abbreviations are as follows: AQ: *Amphimedon queenslandica* (sponge), DM: *Drosophila melanogaster* (fruit fly), HM: *Hydra magnipapillata* (hydra), HS: *Homo sapiens* (human), NV: *Nematostella vectensis* (sea anemone), SP: *Strongylocentrotus purpuratus* (sea urchin).

with our UCR set. Unlike in the fruit fly where 42 out of 64 UCRs overlapped with data reported by Glazov et al. [3], we could not find any UCR in human that overlapped with previously reported UCEs [1] (Additional file 3).

To understand this incongruence, we tested if our pipeline could recover known UCEs of the human-mouse-rat alignment with the same species list and length constraint (≥200 bp) of Bejerano et al. [1]. Our pipeline recovered 464 out of 481 known human UCEs that are conserved both in mouse and rat. The missing 17 known UCEs overlapped with repetitive regions, and these elements could not be recovered by our pipeline, which masks repetitive elements. Furthermore, the human UCEs that were conserved in mouse and rat identified by our pipeline did not also overlap with those newly identified in this study, suggesting that our pipeline works properly. The effect of the genome assembly version used for UCE identification was also negligible as explained above. On the other hand, our stringent repeat masking reduced the number of detectable known UCEs. The numbers of known UCEs were 304, 20,602, and 83 for human-mouse-rat, *D. melanogaster-D. pseudoobscura*, and *D. melanogaster-A. gambiae*, respectively, when we removed known UCEs with simple and known repetitive elements by repeat-masked chromosomes [22], CENSOR [23], and tandem repeat finder [24], the same criteria that we used in this study. However, the most important factor contributing to the identification of novel UCRs was the length constraint (50 bp for human) and species compared. To test this further, our human UCR set was divided into 50 bp sub-sequences, and then a search for these sub-sequences in the genomes of mouse and rat was conducted. Of 28 UCRs, one sub-sequence occurred in both the mouse and rat genomes with 100% similarity. On the other hand, the other 28 UCRs were not conserved in both species, suggesting that those sequences were no longer under strong selective pressure in rodents and could therefore not be identified by the traditional human-mouse-rat alignment (Additional file 3). Indeed, large portions of identified human UCEs are positioned in less conserved loci in placental mammals (Figure 2), which further supports our findings of novel highly conserved DNA elements in model organisms.

UCR clusters arose independently

We then sought evidence for if UCRs from the same or different species share similarity. Considering the short length of UCRs and also assuming that distal regions of ultraconserved elements have higher mutation rates than proximal regions [15,25,26], we analyzed UCRs and their 50 bp-flanking sequences. In all, 4,817 UCRs with flanking sequences from all species were clustered, and orthologous and paralogous UCRs were defined. This yielded 61 clusters, of which the largest cluster consisted of 1,168 UCRs from hydra, sea anemone, and sea urchin (Additional file 4).

Although there are large numbers of UCRs across different taxa, we found that UCRs share sequence similarities and that each cluster of UCRs has a distinct species composition. Moreover, Cnidarian UCRs show a tight association, while human UCRs are largely clustered together with those of sea urchin and/or fruit fly (Additional file 4). Gain of essential functions for the survival of the species in ancestral sequences might contribute to the conservation of the sequence in a specific lineage [14]. Another possible explanation would be that even if the ancestral sequences were not beneficial to the species, random sampling contributed to the elimination of other alleles and the fixation of these sequences in the downsized population, creating a new lineage, due to natural catastrophe or population migration, referred to as a “genetic drift” or “population bottleneck” [27]. Although further study is required to explain the immutability of UCEs after lineage divergence and sequence fixation across a long evolutionary history, we cannot rule out this possibility. It also should be noted that the

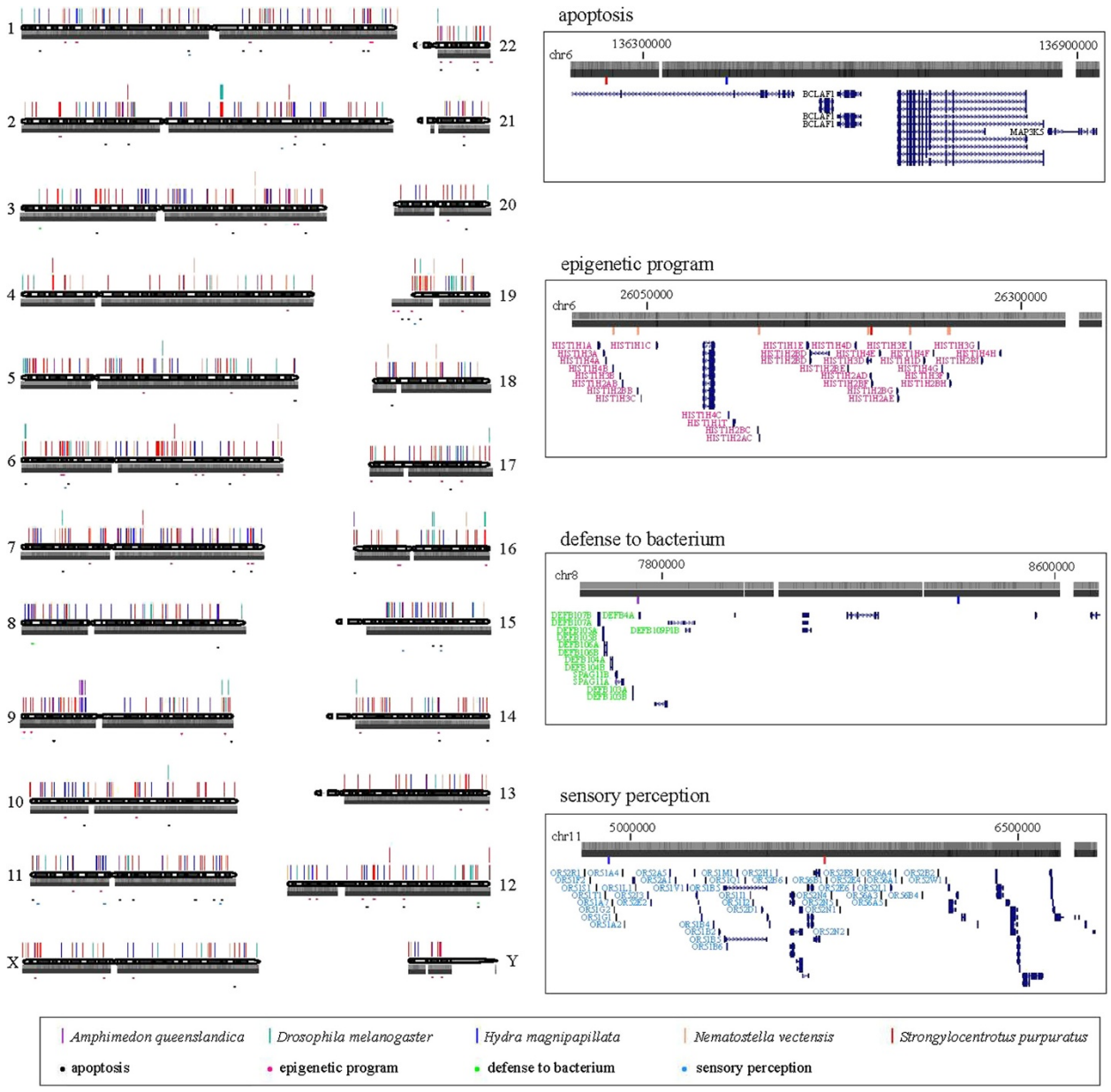

Figure 2 UCEs in the human genome. The short conserved elements (≥30 bp) are depicted above each chromosome and the UCEs (≥50 bp) are depicted above them. The conservation level of human DNA across primates or placental mammals is shown below each chromosome, where the darker color indicates more conservation. Species in which the human sequences are conserved and selected gene categories are labeled as indicated at the bottom. The selected region for each gene category is magnified for a detailed view in the right panel. R package quantsmooth [63] and the UCSC custom track [22] were used for the plot.

absence of UCRs in species from the same lineage does not necessarily mean that those UCRs disappeared in those species but rather that they may exist as derivative sequences by mutation [2,15,28,29].

As shown in Figure 3A and Additional file 5, UCR clusters are clearly separated in a Minimum Curvilinear Embedding (MCE) plot [30], although species is not a good factor to distinguish UCRs (Figure 3B). Short UCRs (≥30 bp) also followed a similar pattern. Interestingly, some clusters have nearly symmetric elements on the MCE plot and it turns out that they are partially reversed complementary sequences.

Network topology demonstrates the relationship between these UCR clusters, where some clusters are connected due to the sequence similarity between components, although most clusters do not share sequence

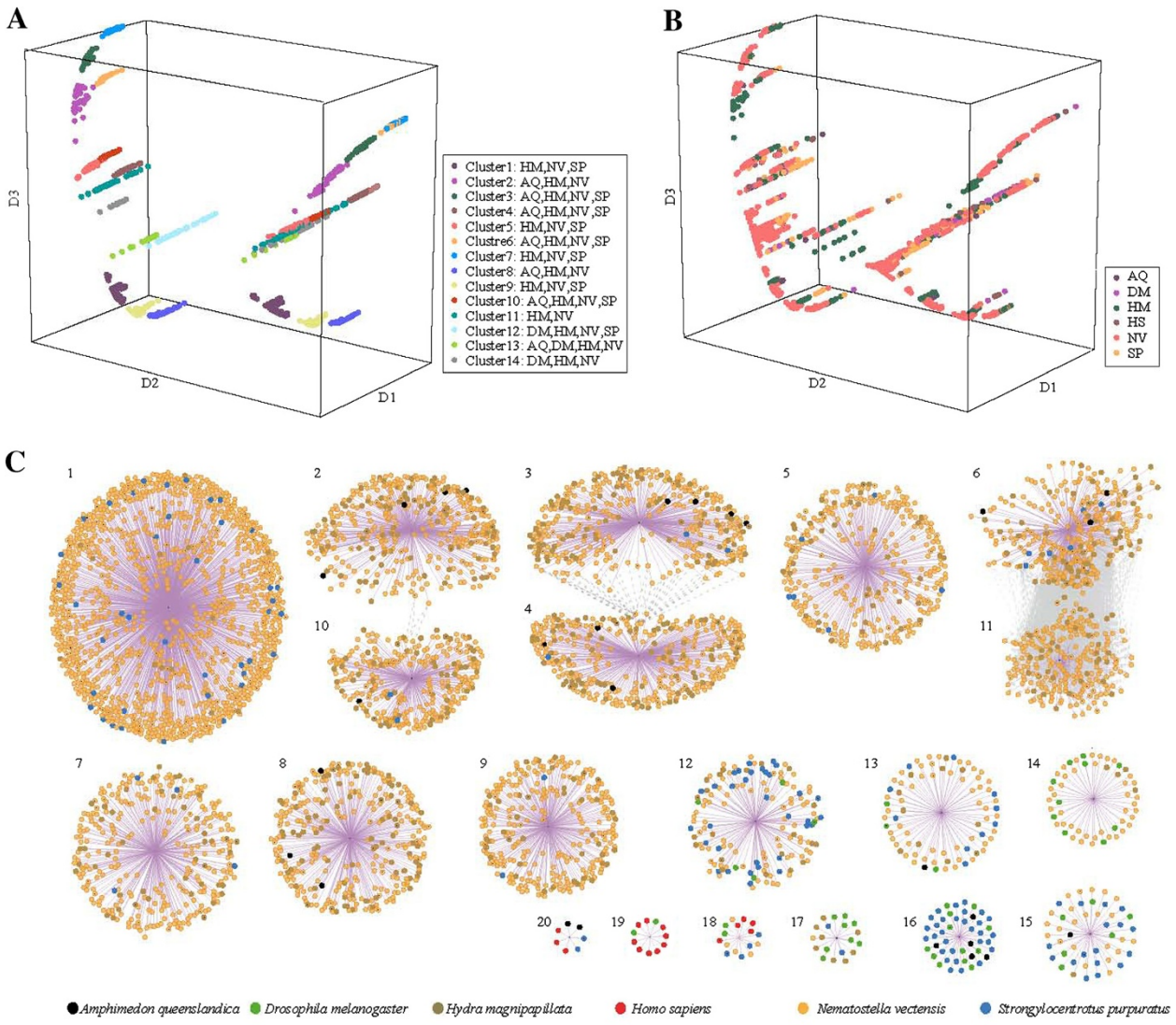

Figure 3 UCR clusters arose independently during evolution. A and **B**. The 5-mer composition of UCRs and 50 bp-flanking sequences were taken as input for the Minimum Curvilinear Embedding analysis and the top three dimensions are depicted here. Clusters and species are marked with different colors as indicated in the inset on the lower right corner. **C**. UCR cluster relationships. Each node represents a UCR with flanking sequence. The similarity between nodes in a same cluster is omitted to avoid extreme density. A cluster centroid is made instead and connected to the components to show membership within the cluster (purple lines). The gray lines show sequence similarities between nodes in different clusters. Clusters with fewer than 7 nodes are not shown. UCRs from sea anemone predominate in this figure due to the large number (3,807 among all 4,817 UCRs).

similarity with others and have unique species composition (Figure 3C). Thus, the UCRs of each cluster may have their own independent origin in a specific lineage.

The neighboring genes of UCRs have distinct functions

UCEs are often flanked by developmental genes, TFs, ion channels, or splicing factors [5,7]. We investigated the functions of each cluster's nearby genes. Due to the paucity of functional annotations of genes and the short length of genome scaffolds in non-model species (Additional file 6), we focused our analysis on the protein domains of nearby genes within 100 kb from UCRs. Neighboring genes to UCR clusters span a spectrum of statistically significant protein domains. However, each cluster is enriched with a distinct set of domains (Figure 4).

Ion channel and transporter domains are the predominant categories; they appear in many clusters composed of various species. Neurotransmitter-gated ion channels and sodium or calcium ion exchanger genes are over-represented in clusters 13, 15, and 17, whose UCRs are

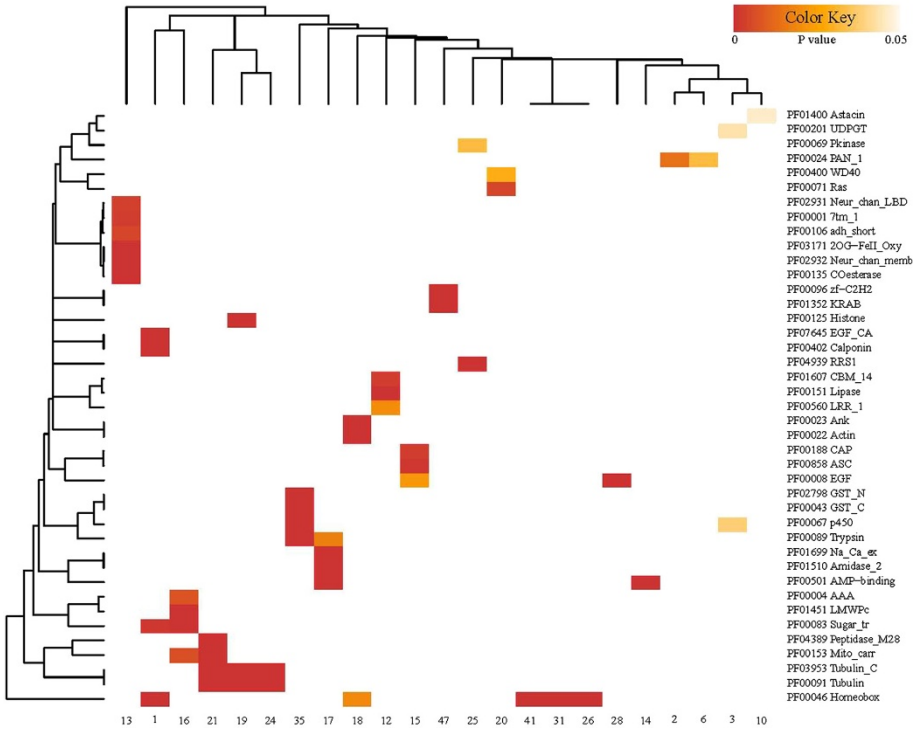

Figure 4 Protein domain enrichment of UCR flanking genes. Association of domains (rows) or clusters (columns) is depicted in dendrograms on the right and upper side of the heatmap, respectively. Only clusters having at least 10 genes were analyzed. Domains whose p-value. <0.05 in at least one cluster and that occurred in at least three nearby genes are shown on the heatmap for visualization purpose.

conserved in all species considered here but human (Figure 4 and Additional file 4). Cation transporters are identified in cluster 30, which consists of human and fruit fly UCRs. Sugar transporters and mitochondrial carrier domains that transport various molecules across membranes are enriched in clusters 1, 16, and 21. These observations are probably because ion channels and transporters are crucial in all living organisms for the maintenance of water, salt, and nutrient homeostasis as well as for electric signal transmission in neuronal and muscle cells [31].

The homeobox domain, part of the TFs that act during the developmental process, is enriched in five clusters. This domain is found in all six species, with three of the five enriched UCR clusters composed of UCRs from human and fruit fly, one from fruit fly and sea urchin, and the last cluster from hydra, sea anemone, and sea urchin. Fruit fly genes regulating developmental programs ranging from axis patterning to molting, such as *bicoid*, *fushi tarazu*, and ecdysone receptor, are also found in several clusters, even those without significant domains.

Histones are overrepresented in cluster 19, which consists of sea anemone and sea urchin UCRs. Evidence that chromatin-related genes flank conserved elements in human (Additional file 7) and from other studies [32,33] suggest that there is a liaison between conserved elements and epigenetic control mechanisms.

Detoxification domains such as cytochrome p450, UDPGT, and GST are enriched in cluster 3 and cluster 35. Cluster 3 consists of UCRs from sponge, hydra, sea anemone, and sea urchin; cluster 35 consists of UCRs from fruit fly and human. These enzymes are important to catalyzing and eliminating endogenous and exogenous substrates and therefore to providing a healthy environment for the cellular system [34]. This remarkable linkage between UCRs and detoxification mechanisms has not previously been reported to our knowledge.

Further analysis of UCRs (≥50 bp) and short UCRs (≥30 bp) in human reveals similar but more interesting properties in terms of nearby gene functions and species conservation (Additional file 7 and Additional file 8). Genes acting in various developmental processes are highly enriched near the UCRs in human that are also conserved in fruit fly and sea urchin. To our surprise and contrary to previous studies, few genes related to development are enriched near the human sequences conserved in sponge, hydra, or sea anemone. Expansion of the relationship between developmental programs and UCRs in human, fruit fly and sea urchin (Figure 1 and Additional file 7 and Additional file 8) implies that the association of conserved sequences with the regulation

of developmental genes started or expanded after the divergence of the Bilateria lineage from the metazoan stem. Our UCR clustering results bolster this hypothesis (Figure 4). Four out of five UCR clusters that have overrepresented homeobox domains of nearby genes come from human, fruit fly, and sea urchin.

Interestingly, genes surrounding short UCRs are enriched with epigenetic program-related genes (Figure 2 and Additional file 7). Short UCRs conserved in human and in fruit fly, hydra, sea anemone, or sea urchin are located near histone gene clusters across several chromosomes. Furthermore, many important epigenetic regulators are also found near elements conserved in sponge, hydra, sea anemone, or sea urchin. These include histone demethylases (KDM3B, KDM4C, KDM5C, and KDM5D), histone acetyltransferases (EP300 and KAT7), histone deacetylases (HDAC2 and HDAC10), retinoblastoma-like protein (RBL1), polycomb ring finger oncogene (BMI1), chromodomain helicase (CHD8), and components of the chromatin remodeling complex, SWI/SNF (SMARCA2, SMARCB1, SMARCC2, and SMARCD3). Taken together with the previously suggested relationship between highly/ultraconserved elements and epigenetic control [15,32,33], our results suggest an interesting hypothesis that epigenetic control mechanisms have tight relationships with conserved DNA sequences and that they might have coevolved from metazoan ancestors rather than recently developed.

Genes implicated in apoptosis, olfactory reception, and defense mechanisms are also enriched near DNA elements conserved in sponge, hydra, or sea urchin (Figure 2 and Additional file 7 and Additional file 8). Our analysis suggests that genomes preserve ancestral sequences well, and these ancestral sequences might have coevolved with a diverse set of essential genes. When and how genes and conserved elements initiated their relationships remains unclear and the mechanism for such an association needs to be further elucidated. However, our analysis expands the repertoire of conserved genomic elements that are possible regulatory elements.

UCRs are enriched with binding sites for developmental TFs

The enhancer activities of UCEs have been reported by several studies [8,9]. To investigate the possibility that these enhancer activities were also conserved in primitive species, we identified significantly overrepresented oligomers and related TF binding sites (TFBSs) for each UCR cluster (Figure 5).

Among 31 TFs that had significant 8-mer matches, 28 were implicated in developmental processes and many were homeobox TFs. Binding sites of homeobox TFs on UCEs near the developmental genes in higher eukaryotes have been identified [35-37], although our clustering results identified various nearby gene categories that were not limited to developmental genes. Prevalent occurrence of developmental TFBSs regardless of cluster and species may be an indication that extensive binding of developmental TFs on UCEs existed in metazoan ancestors and these TFs regulated various nearby genes to coordinate developmental functions. These may have contributed to the strong selective pressure on UCEs that function as regulatory sequences.

Conclusions

Genomes are dynamic entities and are under selective evolutionary pressure from mutation and fixation. Beneficial or neutral mutations in the ancestors of specific lineages are maintained in the population and vertically transferred to descendants [38]. However, these dynamic and selective pressures are not applied uniformly across the whole genome [16,39,40]. Deleterious mutations in essential regions are corrected in a population [15,16]. Sequence conservation thus implies that the function of the sequence is essential. Despite controversy about the indispensability of ultraconserved elements [13,41], much work has demonstrated various vital functions of such elements [5,6,8-10].

As more genomes from various taxa are being sequenced, the opportunity to understand genome conservation and usage increases. Here, we compared genome sequences ranging from primitive aquatic to higher terrestrial species and described for the first time a number of novel UCEs present in primitive species as well as previously uncharacterized UCEs in human and fruit fly. We observed that UCEs cluster by sequence similarity and each cluster has distinct patterns of species composition. These UCEs also exhibited specific biases toward the function of nearby genes and oligomer compositions of the UCE sequences, suggesting that each group of UCEs was generated in the common ancestors of specific lineages and fixed during the evolution of descendants. Although a more detailed functional analysis of UCEs cannot currently be conducted due to the nature of the short draft sequences and because gene functions of non-model species have been less studied, our analysis suggests that UCEs harbor important sequence features, such as binding sites of developmental TFs to coordinate the expression of essential genes, which is why they were readily conserved over the long course of evolution.

Methods

Data preparation

Genome sequences, gene annotation, and protein sequences were downloaded from the UCSC database for human (assembly version: hg19) and fruit fly (assembly

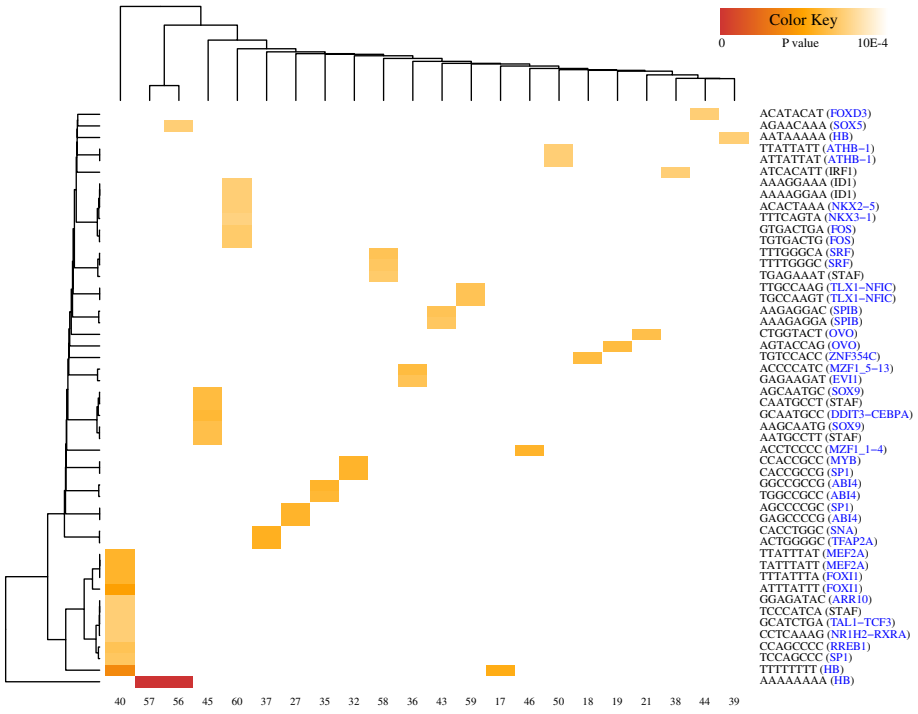

Figure 5 Oligomer enrichment on UCRs. Each cluster (column) shows distinct TFBS patterns. Association of 8-mers and clusters is depicted in dendrograms. Only 8-mers with p-value<1E-4 are shown by the heatmap. The most related TFBSs of each 8-mer are shown in the brackets on the right side. TFs with developmental functions indicated by NCBI Gene [64] or GeneCards [65] are colored blue.

version: dm3), and each genome project for sponge (assembly version as of 5 Aug 2010) [42], hydra (assembly version as of 28 Jan 2009) [43], sea anemone (assembly version as of 26 Oct 2005) [44], and sea urchin (assembly version as of 13 Oct 2006) [45].

Phylogenetic analysis

First, we identified single copy genes from each of six species under investigation to infer their phylogenetic relationships. This approach had been used previously in other studies to avoid the paralogy issue [44,46,47]. Inparanoid was used to identify orthologs and paralogs between species pairs [48]. Only the longest peptide was used when multiple transcripts came from the same gene. We identified 472 single-copy genes that were found to be largely involved in ribosome, spliceosome, or proteasome pathways. Gene sequences were aligned using MUSCLE [49] and the evolutionary distance and phylogenetic tree were obtained using MEGA5 [50]. The phylogenetic tree reveals the overall relationship between six species, which was in agreement with the known classification of these lineages (Figure 1) [45,51,52].

Identification of ultraconserved elements

To identify UCEs for all species pairs, we masked repetitive sequences in the scaffolds of sponge, hydra, sea anemone, and sea urchin using CENSOR [23] and tandem repeats finder [24]. Repeat-masked chromosomes from the UCSC database were used for human and fruit fly [22]. To identify non-gapped conserved elements between two species, we used MUMmer, which rapidly aligned long sequences and detected exact matches using the suffix tree algorithm, with the *maxmatch* option to compute all maximal identical matches regardless of uniqueness [20]. Both forward and reverse complement matches were reported. Identical matches equal to or longer than 50 bp were identified, and ≥30 bp matches were also identified for incidental analysis. Identified UCEs were further masked using CENSOR and tandem repeat finder again. It should be mentioned that this stringent repeat-masking process may have deleted potential UCEs containing repetitive elements.

Two UCEs were joined if they overlapped, and this merging process was repeated until no two UCEs overlapped (Additional file 1 and Additional file 2). Fifty base flanking sequences on both sides of merged UCEs were retrieved using the custom python script.

Clustering of ultraconserved elements

Merged ultraconserved elements with flanking sequences were grouped by sequence similarity. Pairwise alignment of all sequences was computed using BLASTN [53]. The

score density, i.e. the BLAST bit-score divided by the alignment length, was used as the similarity measure. Sequences were clustered using the Markov cluster (MCL) algorithm [54] with default parameters (Additional file 4). In the Minimum Curvilinear Embedding (MCE) analysis [30], 5-mer compositions of the sequences were used as features. In particular, we used the new singular-value-decomposition-based algorithm to implement MCE [55], using the Matlab code provided on the author's website (https://sites.google.com/site/carlovittoriocannistraci/home). The embedding was performed without centering the minimum curvilinear kernel (non-centered MCE).

Nearby genes analysis

Flanking genes within 100 kb of the merged UCEs were obtained from all species under study. For human and fruit fly, we used the gene models from RefSeq [56]. We used the gene models from the respective genome sequencing projects of the non-model metazoans.

Pfam domains of nearby genes were annotated using Interproscan [57] for functional analysis of UCEs. For each domain in each UCR cluster, the domain enrichment of nearby genes within 100 kb of UCRs was calculated using cumulative hypergeometric distribution:

$$\mathrm{P} = \sum_{i=d}^{\min(D,g)} \frac{\binom{D}{i}\binom{G-D}{g-i}}{\binom{G}{g}},$$

where G is the total number of genes from the species pool in the cluster, g is the number of selected nearby genes in the species pool in the cluster, D is the number of occurrences of the domain in the species pool in the cluster, and d is the number of occurrences of the domain in the selected nearby genes in the species pool of the cluster.

Gene ontology enrichment of the nearby genes was analyzed using DAVID [58]. Considering that human has the most comprehensive biological process terms and nearly nothing is annotated in non-model species, only human UCRs and their nearby genes were analyzed.

Motif analysis

A representative sequence of each cluster was generated using MUSCLE [49] and the seqinR package in R [59]. To assess the statistical significance of overrepresented 8-mers, we generated a 10 kb background sequence for each cluster. The background sequence was a combination of segments chosen randomly from all genomes, and each genome contributed to the background with an amount equal to the ratio of its species in the cluster composition. A cumulative binomial probability of observing the given number of the oligomer or more in each cluster was then computed as follows:

$$F(x|n,p) = 1 - \sum_{i=0}^{x-1} \binom{n}{i} p^i (1-p)^{n-i},$$

where x is the number of occurrences of the oligomer, n is the sample size, i.e. sequence length - oligomer size + 1, and p is the probability of observing such an oligomer in the random background sequence. Related TFs for oligomers were identified using STAMP [60].

Additional files

Additional file 1: UCR information (≥50 bp) for each species.

Additional file 2: Short UCR information (≥30 bp) for each species.

Additional file 3: Comparison between identified UCRs and previously known fruit fly UCEs.

Additional file 4: Clustering results of UCRs with 50 bp flanking sequences.

Additional file 5: Visualization of the UCE clusters by minimum cuvilinear embedding.

Additional file 6: Distribution of scaffold length for non-model species.

Additional file 7: Gene ontology enrichment of nearby genes within 50 kb, 100 kb, and 200 kb from short UCRs. Only biological process terms are used.

Additional file 8: Gene ontology enrichment of nearby genes within 50 kb, 100 kb, and 200 kb from UCRs. Only biological process terms are used.

Abbreviations

UCE: ultraconserved element; TF: transcription factor; SNP: single nucleotide polymorphism; AQ: *Amphimedon queenslandica*; DM: *Drosophila melanogaster*; HM: *Hydra magnipapillata*; HS: *Homo sapiens*; NV: *Nematostella vectensis*; SP: *Strongylocentrotus purpuratus*; UCR: ultraconserved region; MCE: Minimum Curvilinear Embedding; TFBS: TF binding site.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TR1 (Taewoo Ryu) and TR2 (Timothy Ravasi) conceived the overall study; TR1 and LS performed the data analysis; TR1, LS, and TR2 drafted the manuscript; all authors read and approved the final manuscript.

Acknowledgements

The authors thank Dr. Carlo Vittorio Cannistraci (KAUST) for conducting the visualization analysis by Minimum Curvilinear Embedding and for the creation of the 3D movie in the supporting information. We also are grateful to Professor Christoph Gehring for his critical review on evolutionary analysis. All the authors are supported by King Abdullah University of Science and Technology.

Received: 26 April 2012 Accepted: 9 November 2012
Published: 5 December 2012

References

1. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome.** *Science (New York, NY)* 2004, **304**(5675):1321–1325.
2. Ovcharenko I: **Widespread ultraconservation divergence in primates.** *Mol Biol Evol* 2008, **25**(8):1668–1676.

3. Glazov EA, Pheasant M, McGraw EA, Bejerano G, Mattick JS: **Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing.** *Genome Res* 2005, **15**(6):800–808.
4. Zheng W-X, Zhang C-T: **Ultraconserved elements between the genomes of the plants Arabidopsis thaliana and rice.** *J Biomol Struct Dyn* 2008, **26**:1–8.
5. Papatsenko D, Kislyuk A, Levine M, Dubchak I: **Conservation patterns in different functional sequence categories of divergent Drosophila species.** *Genomics* 2006, **88**(4):431–442.
6. Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, O'Brien G, Shiue L, Clark TA, Blume JE, Ares M: **Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay.** *Genes Dev* 2007, **21**(6):708–718.
7. Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE: **Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements.** *Nature* 2007, **446**(7138):926–929.
8. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, *et al*: **In vivo enhancer analysis of human conserved non-coding sequences.** *Nature* 2006, **444**(7118):499–502.
9. Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA: **Ultraconservation identifies a small subset of extremely constrained developmental enhancers.** *Nat Genet* 2008, **40**(2):158–160.
10. Derti A, Roth FP, Church GM, Wu C: **Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants.** *Nat Genet* 2006, **38**:1216–1220.
11. Yang R, Frank B, Hemminki K, Bartram CR, Wappenschmidt B, Sutter C, Kiechle M, Bugert P, Schmutzler RK, Arnold N, *et al*: **SNPs in ultraconserved elements and familial breast cancer risk.** *Carcinogenesis* 2008, **29**(2):351–355.
12. Poitras L, Yu M, Lesage-Pelletier C, Macdonald RB, Gagn J-P, Hatch G, Kelly I, Hamilton SP, Rubenstein JLR, Poirier GG, *et al*: **An SNP in an ultraconserved regulatory element affects Dlx5/Dlx6 regulation in the forebrain.** *Development (Cambridge, England)* 2010, **137**(18):3089–3097.
13. Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM: **Deletion of ultraconserved elements yields viable mice.** *PLoS biology* 2007, **5**(9):e234–e234.
14. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D: **A distal enhancer and an ultraconserved exon are derived from a novel retroposon.** *Nature* 2006, **441**(7089):87–90.
15. Stephen S, Pheasant M, Makunin IV, Mattick JS: **Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock.** *Mol Biol Evol* 2008, **25**(2):402–408.
16. Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D: **Human genome ultraconserved elements are ultraselected.** *Science (New York, NY)* 2007, **317**(5840):915–915.
17. Lin Z, Ma H, Nei M: **Ultraconserved coding regions outside the homeobox of mammalian Hox genes.** *BMC Evol Biol* 2008, **8**:260–260.
18. Sakuraba Y, Kimura T, Masuya H, Noguchi H, Sezutsu H, Takahasi KR, Toyoda A, Fukumura R, Murata T, Sakaki Y, *et al*: **Identification and characterization of new long conserved noncoding sequences in vertebrates.** *Mammalian genome: official journal of the International Mammalian Genome Society* 2008, **19**(10–12):703–712.
19. Hedges SB, Dudley J, Kumar S: **TimeTree: a public knowledge-base of divergence times among organisms.** *Bioinformatics (Oxford, England)* 2006, **22**(23):2971–2972.
20. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5**(2):R12–R12.
21. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**(3):R25–R25.
22. UCSC Genome Browser: http://genome.ucsc.edu/.
23. Kohany O, Gentles AJ, Hankus L, Jurka J: **Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor.** *BMC Bioinforma* 2006, **7**:474–474.
24. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27**(2):573–580.
25. Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC: **Itraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales.** *Syst Biol* 2012.
26. Halligan DL, Oliver F, Guthrie J, Stemshorn KC, Harr B, Keightley PD: **Positive and negative selection in murine ultra-conserved noncoding elements.** *Mol Biol Evol* 2011, **28**(9):2651–2660.
27. Gherman A, Chen PE, Teslovich TM, Stankiewicz P, Withers M, Kashuk CS, Chakravarti A, Lupski JR, Cutler DJ, Katsanis N: **Population bottlenecks as a potential major shaping force of human genome architecture.** *PLoS genetics* 2007, **3**(7):e119.
28. Wang J, Lee AP, Kodzius R, Brenner S, Venkatesh B: **Large number of ultraconserved elements were already present in the jawed vertebrate ancestor.** *Mol Biol Evol* 2009, **26**(3):487–490.
29. Kim SY, Pritchard JK: **Adaptive evolution of conserved noncoding elements in mammals.** *PLoS genetics* 2007, **3**(9):1572–1586.
30. Cannistraci CV, Ravasi T, Montevecchi FM, Ideker T, Alessio M: **Nonlinear dimension reduction and clustering by Minimum Curvilinearity unfold neuropathic pain and tissue embryological classes.** *Bioinformatics* 2010, **26**(18):i531–i539.
31. Dubyak GR: **Ion homeostasis, channels, and transporters: an update on cellular mechanisms.** *Adv Physiol Educ* 2004, **28**(1–4):143–154.
32. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K-i, *et al*: **Control of developmental regulators by Polycomb in human embryonic stem cells.** *Cell* 2006, **125**(2):301–313.
33. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, *et al*: **A bivalent chromatin structure marks key developmental genes in embryonic stem cells.** *Cell* 2006, **125**(2):315–326.
34. Ishii Y, Takeda S, Yamada H, Oguri K: **Functional protein-protein interaction of drug metabolizing enzymes.** *Front Biosci* 2005, **10**:887–895.
35. Chiang CWK, Derti A, Schwartz D, Chou MF, Hirschhorn JN, Wu C-T: **Ultraconserved elements: analyses of dosage sensitivity, motifs and boundaries.** *Genetics* 2008, **180**:2277–2293.
36. Lampe X, Samad OA, Guiguen A, Matis C, Remacle S, Picard JJ, Rijli FM, Rezsohazy R: **An ultraconserved Hox-Pbx responsive element resides in the coding sequence of Hoxa2 and is active in rhombomere 4.** *Nucleic Acids Res* 2008, **36**(10):3214–3225.
37. Rödelsperger C, Köhler S, Schulz MH, Manke T, Bauer S, Robinson PN: **Short ultraconserved promoter regions delineate a class of preferentially expressed alternatively spliced transcripts.** *Genomics* 2009, **94**:308–316.
38. Gogarten JP, Townsend JP: **Horizontal gene transfer, genome innovation and evolution.** *Nat Rev Microbiol* 2005, **3**(9):679–687.
39. McLean C, Bejerano G: **Dispensability of mammalian DNA.** *Genome Res* 2008, **18**(11):1743–1751.
40. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**(6822):860–921.
41. Gross L: **Are "ultraconserved" genetic elements really indispensable?** *PLoS biology* 2007, **5**(9):e253–e253.
42. Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T, Richards GS, Conaco C, Dacre M, Hellsten U, *et al*: **The Amphimedon queenslandica genome and the evolution of animal complexity.** *Nature* 2010, **466**(7307):720–726.
43. Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T, Rattei T, Balasubramanian PG, Borman J, Busam D, *et al*: **The dynamic genome of Hydra.** *Nature* 2010, **464**(7288):592–596.
44. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, *et al*: **Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization.** *Science* 2007, **317**(5834):86–94.
45. Sodergren E, Weinstock GM, Davidson EH, Cameron RA, Gibbs RA, Angerer RC, Angerer LM, Arnone MI, Burgess DR, Burke RD, *et al*: **The genome of the sea urchin Strongylocentrotus purpuratus.** *Science (New York, NY)* 2006, **314**(5801):941–952.
46. Roth AC, Gonnet GH, Dessimoz C: **Algorithm of OMA for large-scale orthology inference.** *BMC Bioinforma* 2008, **9**:518.
47. Dessimoz C, Boeckmann B, Roth AC, Gonnet GH: **Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits.** *Nucleic Acids Res* 2006, **34**(11):3309–3316.

48. Ostlund G, Schmitt T, Forslund K, Kostler T, Messina DN, Roopra S, Frings O, Sonnhammer EL: **InParanoid 7: new algorithms and tools for eukaryotic orthology analysis.** *Nucleic Acids Res* 2010, **38:**196–203.
49. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinforma* 2004, **5:**113–113.
50. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods.** *Mol Biol Evol* 2011, **28**(10):2731–2739.
51. Ryu T, Mavromatis CH, Bayer T, Voolstra CR, Ravasi T: **Unexpected complexity of the Reef-Building Coral Acropora millepora transcription factor network.** *BMC Syst Biol* 2011, **5:**58–58.
52. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, *et al*: **Broad phylogenomic sampling improves resolution of the animal tree of life.** *Nature* 2008, **452**(7188):745–749.
53. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403–410.
54. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res* 2002, **30**(7):1575–1584.
55. Cannistraci CV, Lobato GA, Ravasi T: **Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding.** *submitted*.
56. Pruitt KD, Tatusova T, Klimke W, Maglott DR: **NCBI Reference Sequences: current status, policy and new initiatives.** *Nucleic Acids Res* 2009, **37:**32–36.
57. Zdobnov EM, Apweiler R: **InterProScan–an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17**(9):847–848.
58. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4**(1):44–57.
59. Charif D, Lobry JR: **SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis.** *Structural Approaches to Sequence Evolution* 2007, **1:**207–232.
60. Mahony S, Benos PV: **STAMP: a web tool for exploring DNA-binding motif similarities.** *Nucleic Acids Res* 2007, **35:**253–258.
61. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Comput Appl Biosci* 1992, **8**(3):275–282.
62. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**(4):406–425.
63. Oosting J, Eilers P, Menezes R: *Quantsmooth: Quantile smoothing and genomic visualization of array data*. 2009.
64. NCBI Gene: http://www.ncbi.nlm.nih.gov/gene/.
65. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, *et al*: **GeneCards Version 3: the human gene integrator.** *Database (Oxford)* 2010, **020**.

doi:10.1186/1471-2148-12-236
Cite this article as: Ryu *et al.*: The evolution of ultraconserved elements with different phylogenetic origins. *BMC Evolutionary Biology* 2012 **12**:236.