Genome Biology

Genome Biology
Biology for the post-genomic era

This Provisional PDF corresponds to the article as it appeared upon acceptance. Copyedited and fully formatted PDF and full text (HTML) versions will be made available soon.

Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage

Genome Biology 2006, **7**:R56 doi:10.1186/gb-2006-7-7-r56

Remo Sanges (rsanges@tigem.it)
Eva Kalmar (Eva.Kalmar@itg.fzk.de)
Pamela Claudiani (claudiani@tigem.it)
Maria D'Amato (damato@tigem.it)
Ferenc Muller (Ferenc.Mueller@itg.fzk.de)
Elia Stupka (elia@tigem.it)

ISSN 1465-6906

Article type Research

Submission date 26 March 2006

Acceptance date 27 June 2006

Publication date 19 July 2006

Article URL http://genomebiology.com/2006/7/7/R56

This peer-reviewed article was published immediately upon acceptance. It can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in *Genome Biology* are listed in PubMed and archived at PubMed Central.

For information about publishing your research in *Genome Biology* go to

http://genomebiology.com/info/instructions/

© 2006 Sanges *et al.*, licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage

Remo Sanges (rsanges@tigem.it)[1], Eva Kalmar (Eva.Kalmar@itg.fzk.de)[2], Pamela Claudiani (claudiani@tigem.it)[1], Maria D'Amato (damato@tigem.it)[1], Ferenc Muller (Ferenc.Mueller@itg.fzk.de)[2*], Elia Stupka (elia@tigem.it)[1]*

1 Telethon Institute of Genetics and Medicine, Via P.Castellino,111, 80131 Napoli, Italy

2 Institute of Toxicology and Genetics, Forschungzenbrum, Karlsruhe, Postfach 3640, D-76021 Karlsruhe, Germany

* To whom correspondence should be addressed.

RUNNING TITLE: Shuffling of vertebrate regulatory elements

Abstract

Background: All vertebrates share a remarkable degree of similarity in their development as well as in the basic functions of their cells. Despite this, attempts at unearthing genome-wide regulatory elements conserved throughout the vertebrate lineage using BLAST-like approaches have so far detected non-coding conservation in only a few hundred genes, mostly associated with regulation of transcription and development.

Results: We used a unique combination of tools to obtain regional global-local alignments of orthologous loci. This approach takes into account shuffling of regulatory regions likely to occur over evolutionary distances greater than those separating mammalian genomes. This approach revealed one order of magnitude more vertebrate conserved elements than was previously reported in over 2,000 genes, including a high number of genes found in the membrane and extra-cellular regions. Our analysis reveals that 72% of the elements identified have undergone shuffling. We tested the ability of the elements identified to enhance transcription in zebrafish embryos and compared their activity to a set of control fragments. We show that more than 80% of the elements tested were able to significantly enhance transcription, prevalently in a tissue-restricted manner that corresponds to the expression domain of the neighboring gene.

Conclusions: Our work elucidates the importance of shuffling in the detection of cis-regulatory elements. It also elucidates how similarities across the vertebrate lineage, which go well beyond development, can be explained not only within the realm of coding genes but also that of the sequences that ultimately govern their expression.

Background

Enhancers are cis-acting sequences that increase the utilization and/or specificity of eukaryotic promoters, can function in either orientation and often act in a distance and position independent manner [1]. The regulatory logic of enhancers is often conserved throughout vertebrates and their activity relies on sequence modules containing binding sites that are crucial for transcriptional activation. However, recent studies on the cis-regulatory logic of Otx in ascidians pointed out that there can be great plasticity in the arrangement of binding sites within individual functional modules. This degeneracy, combined with the involvement of a few crucial binding sites, is sufficient to explain how the regulatory logic of an enhancer can be retained in the absence of detectable sequence conservation [2]. These observations, together with the fact that we are still far from understanding fully the grammar of transcription factor binding sites and their conservation [3] make it difficult to assess the extent of conservation in vertebrate cis-regulatory elements .

Very little is known about the evolutionary mobility of enhancer and promoter elements within the genome as well as within a specific locus. There are sporadic studies on selected gene families addressing questions related to the mobility of regulatory sequences involving promoter shuffling [4] and enhancer shuffling [5] which describe the gain or loss of individual regulatory elements exchanged between specific genes in a cassette fashion [6]. These studies suggested that a wide variety of different regulatory motifs and mutational mechanisms have operated upon noncoding regions over time. These studies, however, were conducted prior to the advent of large-scale genome

sequencing and thus they were performed on a scale that would not allow the authors to derive more general conclusions on the mobility and shuffling of regulatory elements.

The basic tenet of comparative genomics is that constraint on functional genomic elements has kept their sequence conserved throughout evolution. The completion of the draft sequence of several mammalian genomes has been an important milestone in the search for conserved sequence elements in non-coding DNA. It has been estimated that the proportion of small segments in the mammalian genome that is under purifying selection within intergenic regions is about 5% and that this proportion is much higher than can be explained by protein-coding sequences alone, implying that the genome contains many additional features (such as untranslated regions regulatory elements, non-protein-coding genes, and structural elements) under selection for biological functions [7-11]. In order to tackle this question sequence comparisons across longer evolutionary distances and in particular with the compact *Fugu rubripes* genome have been shown to be useful for dissecting the regulatory grammar of genes much before the advent of genome sequencing [12]. More recently the completion of the draft sequence of several fish genomes has allowed larger scale approaches to detect several regulatory conserved non-coding features.

Several studies have addressed the question of conserved non-coding sequences on a larger scale. A first study on chromosome 21 revealed conserved non-genic sequences (CNGs) identified by using local sequence alignments between the human and mouse genome of high similarity which were shown to be untranscribed [13]. A separate study focusing on sequences with 100% identity revealed the presence of ultra-conserved elements (UCEs) on a genome-wide scale [14] and finally conserved noncoding elements

(CNEs) were found by performing local sequence comparisons between the human and fugu genomes showing enhancer activity in zebrafish co-injection assays [15]. While the CNG study yielded a very large number of elements dispersed across the genome, and bearing no clear relationship to the genes surrounding them, the latter studies (UCEs and CNEs) were almost exclusively associated with genes which have been termed "trans-dev", i.e. involved in developmental processes and/or regulation of transcription.

One of the major drawbacks of current genome-wide studies is that they rely on methods for local alignment, such as BLAST [16] and FASTA [17], which were developed when the bulk of available sequences to be aligned were coding. It has been shown that such algorithms are not as efficient in aligning non-coding sequences [18]. To tackle this issue new algorithms and strategies have been developed in order to search for conserved and/or over-represented motifs from sequence alignments, such as the motif conservation score (MCS) [19], the threaded blockset aligner program (TBA) [20], the regulatory potential score (RP) [21], as well as phastCons elements and scores [22]. However all of the above rely on a blast-like algorithm to produce the initial sequence alignment and are thus subject to some of the sensitivity limitations of this algorithm and do not constitute a major shift in alignment strategy that would model more closely the evolution of regulatory sequences.

Recently two approaches have been published which provide novel alignment strategies: the promoterwise algorithm coupled with "evolutionary selex" [23], and the CHAOS alignment program [24]. While the former has been used to validate a set of short motifs, which have been shown to be of functional importance, the latter has not been coupled to experimental verification to estimate its potential for the discovery of

conserved regulatory sequences. Unlike other fast algorithms for genomic alignment, CHAOS does not depend on long exact matches, it does not require extensive ungapped homology, and it does allow for mismatches within alignment seeds, all of which are important when comparing noncoding regions across distantly related organisms. Thus CHAOS could be a suitable method for the identification of short conserved regions that have remained functional despite having changed their location during vertebrate evolution. The only method available which attempts to tackle the question of shuffled elements and which makes use of CHAOS is Shuffle-Lagan [25], however it has not been used on a genome-scale and its ability to detect enhancers has not been verified experimentally.

Until recently our ability to verify the function of sequence elements on a large scale within an in vivo context was strongly limited. This task has been eased significantly using co-injection experiments in zebrafish embryos [26], which allows significant scale-up in the quantity of regulatory elements tested, fundamental when trying to understand general principles regarding regulatory elements whose grammar still eludes us. The co-injection technique used to test shuffled conserved regions (SCEs) for enhancer activity was shown previously to be a simple way of testing cis-acting regulatory elements [15, 27, 28] and was shown to be an efficient way to test many elements in a relatively short time [15].

The analysis herein described attempts to tackle the question of the extent, mobility and function of conserved noncoding elements in across vertebrates orthologous loci using a unique combination of tools aimed at identifying global-local regionally conserved elements. We first use orthologous loci from four mammalian genomes to

extract “regionally conserved elements” (rCNE) using MLAGAN [29], and then use CHAOS to verity the extent of conservation of those rCNEs within their orthologous loci within fish genomes. The analysis is conducted annotating the extent of shuffling undergone by the elements identified. Finally we investigate the activity of rearranged and shuffled elements as enhancer elements in vivo. We show that the inclusion of additional genomes, the use of a combined global-local strategy, and the deployment of a sensitive alignment algorithm such as CHAOS yields an order of magnitude increase in the number of potentially functional noncoding elements detected to be conserved across vertebrates and that the majority of these have undergone shuffling and are likely to act as enhancers in vivo based on the >80% rate of functional and tissue-restricted enhancers detected in our zebrafish co-injection study.

Results

Identification of mammalian rCNEs

For each group of orthologous genes global multiple alignments among the human, mouse, rat and dog loci were performed using MLAGAN [25]. We took into consideration all genes for which there were predicted othologs within Ensembl [30] in the mouse genome, human genome and any third mammalian species, which led us to analyze 9,749 groups of orthologous genes (i.e. 36% of the annotated mouse genes). Most genes (~88%) were found conserved in all four species taken into account, with only ~12% found only in 3 out of 4 species (~6% in each triplet, see Figure 1). For each locus we took into account the whole genomic repeat-masked sequence containing the transcriptional unit as well as the complete flanking sequences up to the preceding and following gene. This lead us to analyze overall 37% of the murine genome sequence. The

alignments were parsed using VISTA [31] searching for segments of minimum 100 bp length and 70% identity. We further selected these regions by only taking into account regions that were found at least in mouse, human and a third mammalian species and which overlapped by at least 50bp, which resulted in a set of 364,358 rCNEs (see Table 1). These were then filtered stringently to distinguish "genic" from "non-genic" (see methods). This analysis classified 22.7% of the resulting rCNEs as "genic", and 281,644 non-genic elements making up ~46Mb, i.e. 1.77% of the murine genome.

We further annotated mammalian rCNEs based on their position in the mouse genome with respect to the gene locus in order to define whether they were located prior to the annotated transcription start site ("pre-gene"), within the intronic portion of the gene, or posterior to the transcriptional unit ("post-gene"). Approximately 54% of rCNEs fall within intergenic regions, of which 37% post-gene and 63% pre-gene (see Table 1).

Shuffling of conserved elements is a widespread phenomenon

We searched for conservation of rCNEs in teleost genomes using CHAOS [24] and selecting regions which presented at least 60% identity over a minimum length of 40bp as compared to the mouse sequence of the rCNEs. This method allowed us to identify regions that are reversed or moved in the fish locus with respect to the corresponding mammalian locus. For each locus in every species analyzed we took into account the whole genomic repeat-masked sequence containing the transcriptional unit as well as the complete flanking sequences up to the preceding and following gene. We defined as shuffled conserved elements (SCEs) regions of the mouse genome conserved at least in the fugu orthologous locus and filtered out any sequence shorter than 20bp as a result of the overlap analysis with zebrafish and tetraodon (see Materials and Methods

for details). Our analysis identified 21,427 non-redundant non-genic SCEs found in ~30% of the genes analyzed (2,911) (see Table 2). The distribution of their length and percentage identity is shown in Figure 2E and 2F. The median length and percentage identity (45bp, 67% respectively) reflect closely the cut-offs provided to CHAOS in the alignment (40bp, 60% identity), although there is a significant number of outliers whose length is equal to or greater than 200bp (223 elements whose maximum length is 669bp), whose median percentage identity is 74%. No elements were identified which were completely identical to their mouse counterpart (the maximum percentage identity found was 97%).

We decided to investigate further to what extent the elements identified, which are still retained within the locus analyzed, have shuffled in terms of relative position and orientation relative to the transcriptional unit and would thus be missed by a simple regional global alignment (e.g. MLAGAN). The results of this analysis show that only 28% of elements identified have retained the same orientation and the same position with respect to the transcriptional unit taken into account (i.e. have remained pre-gene, intronic, or post-gene labeled as "collinear", see Figure 2A), while others have shifted in terms of orientation ("reversed", see Figure 2B), position ("moved", Figure 2C), or both ("moved-reversed", Figure 2D). Thus almost 2/3 of the SCEs identified would have been missed by a global, albeit regional, alignment approach. A possible explanation for the large number of non-collinear elements is that they could appear shuffled owing to assembly artifacts. In order to assess whether the large number of elements identified as non-collinear are merely due to assembly artifacts we analyzed the number of SCEs containing a single hit in fugu and not classified as collinear that also had a match in

tetraodon. If the shuffling was merely due to assembly artifacts we would expect that approximately half of the non-collinear hits in fugu were also found to be non-collinear in Tetraodon. The results, however, were significantly different, since more than 80% of the elements were not collinear in both species (p-value < 2.2e-16 obtained by performing a chi-square comparison between the proportion obtained and the expected 0.5/0.5 proportion). These results underscore that shuffling is a mechanism of particular relevance when searching for short, well-conserved elements across long evolutionary distances and that its true extent can only be detected by using a sensitive global-local alignment approach, rather than a fast genome-wide approach [25].

Two examples of SCEs that were identified in our study are shown in Figure 3: Example A shows the locus of Sema6d, a semaphorin gene located in the plasma membrane and involved in cardiac morphogenesis. This locus presents a conserved element which is found after the transcriptional unit at the 3' end of the gene in all mammals analyzed, while it is located upstream in fish genomes, and reversed in orientation in the fugu and tetraodon genomes. Example B shows the locus of the tyrosine phosphatase receptor type G protein, a candidate tumor suppressor gene, which presents a conserved element in the first intron of all mammalian loci analyzed, which is found in reversed orientation in all fish genomes, downstream of the gene in the fugu and tetraodon genomes, and in the second intron in the zebrafish genome.

SCEs cast a wider net of non-genic conservation across the genome

We analyzed the type of genes that are associated with SCEs by assessing the distribution of Gene Ontology (GO) terms [32] using GOstat [33] (see Materials and Methods). Although the results indicate significant over-representation of gene classes typical of

genes harboring non-coding conservation (e.g. “trans-dev” enrichment) as reported before (see Table S1 for more details), the number of genes within our analysis containing non-genic SCEs (2,911) is approximately an order of magnitude greater than that of the number of genes containing CNEs (330). The overlap between the two datasets is of 291 genes, thus almost all (>88%) genes containing SCEs also contain CNEs. A gene ontology analysis comparing genes containing CNEs and those containing SCEs (see Figure 4) reveals that there are several gene ontology categories which are significantly under-represented in the CNE dataset as compared to ours. These categories are not seen in the previous analysis (seen in Table S1) because they are not over-represented in our dataset as compared to the entire genome.

The most striking difference is found in the analysis by cellular components: there is a ~54-fold enrichment in genes belonging to the extracellular regions which contain SCEs as compared to genes in the same class which contain CNEs. In fact SCEs are present in over 50% of the genes we were able to classify as belonging to the extracellular matrix and in 35% of those belonging to the extracellular space, while CNEs are only found in 6 and 2 such genes respectively. These gene sets differ significantly in both extracellular regions and membrane GO cellular component categories (p-value < 0.001, see Table S1). Enrichments in the order of 10- to 13-fold are seen when comparing genes involved in physiological and cellular processes respectively. For both of these categories our analysis is able to identify SCEs in more than 30% of the genes belonging to this class. The differences, though substantial (~7-fold), are not as extreme when comparing “trans-dev” genes (i.e. genes categorized as belonging to regulation of biological process and development using GO) because the CNE dataset has a stronger bias for those genes (p-

value < 0.001, see Table S1). Finally while we identify SCEs in 40% of genes assigned to the "behaviour" class, none of the genes in this class present CNEs. The data thus suggests that there are both quantitative and qualitative differences between the two datasets.

The proximal promoter region is a shuffling "oasis"

Since a large proportion of our dataset undergoes shuffling we decided to investigate whether shuffling is a property that is dependent on the proximity to the transcriptional unit. To address this question we divided our dataset of non-genic SCEs between collinear (as discussed above) and noncollinear (all other categories discussed earlier taken together) elements, and analyzed the distribution of their distances from the TSS (pre-gene set), the intron start (intron-start), the intron end (intron-end set) and the 3' end of the transcript (post-gene). This analysis demonstrated that collinear elements distributed significantly closer to the start and the end of the transcriptional unit compared to non-collinear elements, while no differences were observed in terms of proximity to the intron start and intron end (see supplementary Figure S1).

In order to investigate this phenomenon at a higher resolution, we subdivided all loci analyzed in our dataset in 1000bp windows within the areas, and verified whether in any of these windows the proportion of collinear vs non-collinear elements deviated significantly from the expected proportions (see methods for details). The results of the analysis are shown in Figure 5. The only window that showed a high chi-square result with significantly less shuffled elements than collinear ones (p-value = e-08), was the 1000bp window immediately upstream of the TSS. No similar results were found in any other 1000bp window across the gene loci analyzed. Similar results were obtained when

deploying other window sizes (data not shown). To ascertain whether the result observed was due to annotation problems we inspected the gene ontology classification of the genes that presented non-genic collinear elements in the 1000bp window discussed above and observed a significant enrichment (p-value < 0.001) for "trans-dev" genes, while the same test conducted on genic collinear elements in the same window revealed no significant gene ontology enrichment (see Table S3).

SCEs are able to predict vertebrate enhancers

In order to verify the ability of SCEs to predict functional enhancer elements we conducted an overlap analysis (see Materials and Methods) of SCEs with 98 mouse enhancer elements deposited in Genbank. We compared the overlap of SCEs with that of two other datasets that present conservation in fish genomes, namely CNEs and UCEs. The results in Figure 6 show that while CNEs and UCEs are able to detect only 1 and 2 known enhancers from our dataset respectively, while SCEs detect 18 of them successfully.

SCEs act as enhancers in-vivo

In order to validate the cis-regulatory activity of SCEs we chose a subset of SCEs to be tested for in vivo enhancer activity by amplifying them from the fugu genome and co-injecting them in zebrafish embryos with a minimal promoter-reporter construct yielding transient transgenic zebrafish embryos. A total number of 27 SCEs was tested of which 4 overlapped known mouse enhancers for which activity had not been previously reported in fish and the remaining 23 (from 12 genes, of which 4 were not trans-dev genes, for a total of 8 fragments not associated to trans-dev genes) did not overlap any known feature. Detailed information on each SCE tested, including diagrams of their localization in mammalian and fish genomes, as well as multiple alignments is shown in supplementary

file S1. As a control set 12 non-coding, non-repeated and non-conserved fragments were also chosen for co-injection assays, of which 9 were from the same genes from which SCEs had been picked and 3 were from random genes (see methods for details). Owing to the mosaic expression patterns which are obtained with this technique, results were recorded in two ways: by counting the number of cells stained for X-Gal and recording where possible the tissue in which the LacZ positive cells were found, as well as by plotting LacZ positive cells on expression maps which represent a composite overview of the LacZ positive cells of all the embryos tested. Results of the cell counts are shown in Table 3 (more details shown in Table S3) and the expression maps are shown in Figure 7. The cell counts were used to define statistically which fragments showed tissue-restricted enhancer activity or generalized enhancer activity (see Materials and Methods).

As a positive control a published regulatory element from the *shh* locus, *ar-C* [27], was coinjected with the *HSP:lacZ* fragment. From a total of 27 SCEs, 22 (i.e. ~81%) were able to significantly enhance the activity of the *HSP:lacZ* construct in comparison to the embryos injected with *HSP:lacZ* only (see methods for details). Of these, 3 out of the 4 tested known mouse enhancers which were found to be conserved in fish were confirmed to act as enhancers in fish. A similar percentage of positive results (82.6%) is obtained if we do not include these enhancers in the count. The enhancer effect in 20 out of the 22 positive SCEs was not generalized but observed in a tissue-restricted manner. The expression patterns obtained in our experiments were compared to expression data retrieved from the Zebrafish Information Network [34, 35]. Multiple SCEs found within a single gene locus gave similar tissue-restricted enhancer activity. For example, all 4 SCEs tested from the *ets-1* locus gave expression that was highly specific to the blood

precursors, (SCE 1646 in Figure 7C). This result is in line with reported data which showed *ets-1* expression in the arterial system and venous system. Moreover both elements tested from the *zfpm2* (also described as *fog2*[36].) gene gave CNS specific enhancer activity, which is in line with a recent report which showed that the expression of both *fog2* paralogs is restricted to the brain[36]. Similarly, elements tested from the *mab-21-like* genes gave CNS and eye specific enhancer activity (SCE 4939, Figure 7F). This pattern of expression corresponds with the patterns reported in the brain, neurons, and eye [37, 38]. The SCEs that were found in the *pax6a* and *hmx3* genes were shown to give CNS specific enhancement, in line with the reported expression of these genes in the CNS [34]. Finally, SCE 3121 from the gene *jag1b* gave specific expression in the CNS and in the eye (Figure 7D) in partial agreement with reported expression of this gene (expressed in the rostral end of the pronephric duct, nephron primordia, and the region extending from the otic vesicle to the eye [39]). Novel enhancer functions were also detected for SCEs neighbouring *lmx1b1*, which showed CNS specific activity, and SCEs neighbouring 4 genes not belonging to the trans-dev category, such as *mapkap1* (Figure 7E)*, tmeff2 and 3110004L20Rik (producing proteins integral to the membrane) and elmo1 (associated to the cytoskeleton), which* showed strong generalized and/or tissue specific activity. No endogenous expression data is available for these genes for comparison. In contrast to the results with SCE elements, only 2 out of 12 (~17%) of the genomic control fragment set derived from the same loci of the SCEs showed significant enhancement of LacZ activity (see Table 3). Taken together, these data demonstrates that SCEs act as bona fide enhancers that can drive tissue-restricted as well as generalized expression during embryo development.

Discussion

Widespread shuffling of cis-regulatory elements in vertebrates

In this study we demonstrate, using a unique combination of tools aimed at obtaining regional, global-local sensitive alignments applied on genome level, that the number of conserved non-genic sequences shared between mammalian and fish genomes is at least an order of magnitude higher than previously proposed and is spread across thousands of genes. In fact, approximately 30% of the genes analyzed presented at least 1 SCE. Our GO analysis results indicate a "trans-dev" bias similar to those described in previous studies addressing genes presenting non-coding conservation [14, 15]. On the other hand, the significant increase in the sheer number of elements identified and in the number of genes presenting SCEs enabled us to detect conserved non-genic elements in a third of the genes studied, indicating that conservation of cis-regulatory modules is a widespread phenomenon in vertebrates, not limited to a few hundred genes as suggested by previous studies . The GO analysis also revealed that certain classes of genes such as those located in the extracellular space and extracellular matrix present conserved non-coding elements which were not identified with previous approaches and indicate that non-coding elements conserved across vertebrates are present in a larger and more diverse set of genes than was previously thought. Although we also see a larger number of genes involved in cellular and physiological processes many of them are are also assigned to "trans-dev" categories and thus their involvement in development and regulation of transcription cannot be excluded. Indeed it is important to point out that 8 out of the 23 randomly selected fragments were not associated to trans-dev genes by GO classification, and 6 of these fragments showed significant enhancer activity in our co-injection assays (table 3) confirming that conservation is not an exclusive characteristic

of regulatory regions associated to trans-dev genes.

Shuffling plays an important role in the fact that we were able to detect such a large number of conserved sequences, as 72% of our dataset has been found either inverted, or moved, or both in the fish locus with respect to the mouse locus. Assembly artifacts are unlikely to be an important factor in the elements identified as shuffled since they would also affect gene structures and therefore correct gene prediction and orthologs detection, which is at the basis of our dataset. We were reassured to this effect by our tetraodon-fugu comparison, which indicated that most elements found as shuffled in one species are also shuffled in the other species. A notable exception to the general shuffling bias in the elements found was a 1000bp window immediately upstream of the TSS. Taking into account that the proximal promoter region is considered to be approximately -250bp to +100bp from the TSS [40], and assuming that TSS annotations in the mouse genes analyzed are precise, this finding suggests that there is a class of enhancer elements which are more constrained in both position and orientation, perhaps working in tight connection to the promoter complex. The fact that the genes containing non-genic collinear elements in this window show the "trans-dev" bias associated with our overall SCE dataset as well as with previous analyses of non-coding conservation reassures us that this result is not a mere product of bad annotation of the first exon in these genes. It is particularly reassuring that performing the same analysis on SCEs found in the same window, but classified as "genic" (and thus more likely to be real evidence of annotation problems) do not present this bias.

Lack of conservation can also be due to the fact that the evolution of regulatory motifs involves constant de novo creation and destruction of them over time due to their

short sequences and plastic nature ([41] reviewed in[42]). The dissection of cis-regulatory elements from different species, however, indicates clearly that there are cases in which although the same transcription factors are involved in the regulation of gene, all sequences that are not responsible directly for the binding of transcription factors are not preserved and thus overall sequence conservation is very poor [2]. Thus the quest for the identification of regulatory conservation must be complemented by a more thorough understanding of the inherent grammar of regulatory sequences which would lead to improved alignment models specifically tailored to regulatory sequences [23].

Conservation vs. function

In the last few years several strategies have been deployed to perform genome-wide sequence comparisons, which in turn identified several novel functional elements in vertebrate genomes, however they have not yet defined how far conservation of non-coding elements can be pushed to identify efficiently functional elements. The approach used to build our dataset is significantly different from previous approaches, because on the one hand it is stringent by focusing on fish-mammal comparisons and on the other hand it is more sensitive than previous approaches because of its CHAOS based alignments and lower length cut-offs. The requirement for conservation in fish genomes in the SCE dataset would thus lead to the loss of mammalian-specific enhancers, but on the other hand is likely to act as a stringent filter for slowly evolving DNA that may be free of any functional constraints. The differences between the SCE dataset and previously published datasets became evident by performing an overlap analysis amongst them (see Figure S2 and methods for details). The partial overlap between the analyzed datasets highlights yet again that the approach used to determine conserved non-genic elements has a notable impact on the elements identified. Approximately 50% of SCEs

do not overlap any known feature, suggesting that the use of non-exact seeds for the initial local alignments has a significant impact on the analysis of non-coding DNA harboring short, well conserved elements and that our dataset is substantially different from previous datasets both quantitatively, and qualitatively.

Ultra Conserved Elements (UCEs) were detected using a whole genome local alignment strategy between human and mouse (though they are often conserved also in fish genomes) and selected for being 100% identical over at least 200 bps [14]. They were shown to be often located in clusters in the proximity of "trans-dev" genes. Poulin et al. showed that the ultraconserved Dc2 element is necessary and sufficient for brain tissue enhancer activity [43] and an ongoing systematic study using transgenic mice has shown enhancer activity for over 60% of the elements tested so far (64] (Pennachio et al. unpublished data). Our dataset overlaps only 45% of the UCE elements, because of its "regional approach" which will miss any elements which are conserved across non-orthologous loci or wich are found beyond the region we took in consideration (i.e. beyond the previous or next gene). Nonetheless the results of our study would indicate that the enhancer function that has so far been associated with them does not explain fully their level of conservation, since our dataset, although rich in enhancers, presents much lower levels of sequence identity and length as compared to UCEs. Only one of the fragments that we tested (SCE 1973 from the *mapkap1* gene) overlaps with a UCE element. The overlap is only 33bps, and there is no further identity with the UCE in fugu, but the element acted as a tissue-restricted enhancer in-vivo nonetheless. A region adjacent to the UCE in mouse (SCE 1973), though not ultra-conserved, is also conserved in fish and acted as a generic enhancer in our assays highlighting the complexity of these

regions and adding to the ongoing debate regarding their function and evolution [44].

A large set of sequences, defined as conserved non-genic sequences (CNGs), was constructed by using pairwise local sequence comparison between the human and mouse genome on Chromosome 21 (identity >= 70%, length >= 100bp), and it was shown that 2/3 of them lacked transcriptional evidence in vivo [13]. The conservation of these regions in other mammalian genomes was later also confirmed [8], however so far they have not been shown to represent functional regulatory elements to a satisfactory scale, thus the specificity of this method in the identification of enhancers is not known. The overlap analysis highlights that although CNGs are three orders of magnitude larger than UCEs and CNEs and they contain the former fully and 96% of the latter, they only overlap approximately half of the SCE dataset. This would suggest that there are qualitative differences between CNGs and our own dataset. Interestingly, it has been shown that megabase deletions of 2 gene deserts containing thousands of CNGs in mice showed no phenotypic effects [45]. The authors state that none of the CNGs contained are conserved in fish and when we inspected these regions we discovered only a single SCE, very close to the boundary of the deletion.

A recent genome wide study of functional noncoding elements conserved in fish genomes used pairwise local sequence comparison between the human and fugu genomes to define 1,400 higly conserved noncoding elements (CNEs) (length >= 100) and found that these were principally associated with developmental genes. Our dataset overlaps only 51% of the CNEs within the loci analyzed, probably because of the regional approach taken which disregards elements conserved across non-orthologous loci. On the other hand more than 88% of the genes which contain CNEs also present SCEs, thus

identifying regulatory elements in the majority of those genes nonetheless. A group of CNEs were shown to act as enhancers when tested in vivo in zebrafish by co-injecting them with promoter/reporter constructs. Our data, compared to the CNE dataset is a radical extension (of an order of magnitude) of similar conserved elements indicating a significant quantitative difference. There is also a qualitative difference, however, as we identify elements in a very broad range of genes, including genes from the extracellular regions and membrane and many genes participating in physiological and cellular processes which are not transcription factors. The quantitative and qualitative differences in our dataset constitute a major departure from previously published datasets which present conservation across vertebrates and clear evidence to be involved in enhancing gene expression, namely CNEs and UCEs.

Thus the lack of overlap between the datasets taken into consideration is probably a compounded effect of methodological differences (e.g. CNEs, versus SCEs), real biological differences (CNGs versus others) and a compound effect of the two differences (UCEs vs CNEs and SCEs). Our results suggest that a large portion of the non-coding genome is composed of enhancers. Although certainly conserved non-coding regions play other roles which were not able to verify, either they constitute a minority, or they are able to perform several functions besides that of enhancers.

Comparative genomics has been applied successfully to the study of regulatory elements in the past, using approaches based on motif libraries. Xie et al. [19] aligned the promoter and 3'UTR sequences from four mammalian genomes by using BlastZ using a regional approach and were able to identify motifs that were over-represented in conserved regions around genes. They showed that these motifs are non-randomly

distributed with respect to gene expression data but did not identify specific instances of the motif as active copies in the genome. Thus this study, besides using a different methodology, focused on mammalian genomes only, as compared to our vertebrate-wide approach and focused on proximal 5' and 3' UTR sequences, discarding introns as a negative control set, assuming that they contain few regulatory elements. Our study was based on sequence alignment, focused on a broader dataset comprising several vertebrate genomes and made use of the full intergenic and intronic sequence for each locus taken in consideration.

Ettwiller et al. [23] propose a novel computational method that also makes use of comparative genomics. Firstly they developed a novel alignment routine, called promoterwise, that models more closely promoter evolution. Then they used an efficient method to allow direct enumeration of all possible motifs up to 12-mers, including motifs with wild cards. Finally active instances of the motif set thus generated were confirmed by searching them in regions that were found to be conserved in the alignment routine. This work was aimed at comparing distantly related genomes, by searching for over-representation in related orthologs across mammalian and fish genomes to identify specific instances of these motifs. Moreover they proved using experiments in Medaka that these active motifs are necessary to drive expression in vivo. This study resembles more closely our strategy as it involves a vertebrate-wide comparison, although it focused only on 5KB promoter sequences.

Motif library based approaches are complementary to our alignment focused approach. One important difference between these approaches is that the computational requirements of motif-based approaches are very high, thus it is not feasible to execute a

motif library approach over a third of the genome sequence, as was done in this work. On the other hand motif library approaches are able to pinpoint specific motifs that are at the core of the regulatory grammar, while our approach uncovers a dataset which is likely to contain a redundant set of regulatory motifs. It would be a natural extenstion of our work to compare these datasets in order to understand shuffling to what extent enhancers can be represented as clusters of simpler motifs as well as to investigate shuffling of enhancers in relation to the shuffling of single motifs.

Towards improved detection of cis-regulatory elements

The fact that despite an increase of an order of magnitude in our dataset, a similar ratio of elements was found to act as enhancers as compared to the CNE dataset suggests that the extent of sequence conservation of regulatory elements is a moving target that reflects the technique used to identify them. There is a clear need for novel methodologies to detect so far hidden conserved elements. The algorithm Shuffle-LAGAN is an alignment program that resembles our approach although it only aligns shuffled elements within pairwise alignments and therefore it would have not helped to bypass the initial step of selecting rCNEs found conserved in at least 3 mammalian genomes. A desirable extension of Shuffle-Lagan would be to add the ability to process orthologous loci from several genomes at once. More knowledge about the evolution of non-coding DNA will be needed in order to obtain better scoring schemes and thus yield not only sensitive alignments but more reliable predictions of enhancers and other regulators of gene expression [25].

An important aspect that differentiates our approach from previous BLAST-based approaches is the use of CHAOS for alignment of mammalian loci to fish loci. In order to

verify to what extent CHAOS differs from BLAST in this particular type of search we performed the search for SCEs from our set of rCNEs in the fugu genome comparing NCBI BLAST and CHAOS at different word sizes and identical length and identity cut-offs. The results indicate that while CHAOS scales exponentially as word size decreases, the number of hits obtained with BLAST is almost unaltered by the difference in word size. Moreover there is a qualitative difference in the hits obtained since the increase in number of elements identified at small word sizes using CHAOS is due in great part to shuffled elements that BLAST is unable to identify (see Figure S3 for details). This qualitative difference is most notable using word size 10, where only ~4% of BLAST results are shuffled elements as compared to 72% of the elements identified by CHAOS.

This significant difference reiterates quite clearly that looking for sequence similarity across long stretches of identical words is not a valid approach in identifying conserved regulatory elements. At the same time if we were to decrease word sizes to what would be biologically sensible (e.g. word size 5-8, similar to the size of transcription factor binding sites) it would be difficult to assess whether the elements identified as conserved were the result of convergent transcription factor binding site architecture generated de novo, rather than truly conserved across vertebrate evolution. Thus novel methodologies need to be developed which would make use of small word sizes but include other constraints and scoring systems which would help to distinguish biological features preserved through evolution from neutrally evolving short fragments in the genome. To this extent a well curated resource collecting known enhancers (deposited in GenBank for example) as well as a large set of systematically validated

enhancers (e.g Enhancer Browser [46], L.A.Pennacchio unpublished) would help in building valid scoring systems and improve current methods.

In-vivo transient assays

Our in-vivo assays by co-injection revealed interestingly that most enhancers idenfied using this method were restricted in their activity to one or two tissues. Reassuringly the expression profile of 24h old embryos co-injected with the *ArC* positive control showed clear notochord enhancement (Figure 7B) as described previously [27]. The relative evolutionary closeness of fugu and zebrafish implies that expression and regulation of expression of developmentally regulated genes is likely well conserved [15, 47]. Very little is known about Fugu gene expression patterns, but the availability of gene expression pattern information for many zebrafish genes provides a reliable assessment for the tissue specificity of the Fugu SCEs tested in our transient transgenic embryo assays. The functional analysis of SCEs by enhancer essays carried out in the transient transgenic zebrafish identified several new tissue restricted enhancer functions for genes where the endogenous expression pattern is not known. Future work will be required to analyse the role of these enhancers in relation to the detailed analysis of expression patterns of the genes they are associated to. In several cases the SCEs found within a locus provided tissue specificity reminiscent of the gene expression pattern of the flanking gene, arguing strongly for a direct role of these SCEs in regulating the expression of the flanking gene. It will, however, only be possible to unequivocally prove the requirement of these enhancers for driving the expression of the candidate gene by site specific mutation of the SCEs in the genomic context. Two of the control fragments which do not contain detectable conservation were also shown to have significant

enhancer effect and in particular one of the two presented activity that was higher than that of most SCEs tested.

Mechanisms for genome-wide shuffling

Genomic rearrangements had already been reported on a large scale when looking at gene order in regions of synteny between human and Fugu [48]. Similar rearrangements should be seen when analyzing smaller regulatory regions which could harbor enhancers, which present strong evolutionary constraints on their sequence, but often not on their specific localization with respect to the gene they act upon. We show that shuffling and rearrangements are not only applicable to non-genic sequences, but are also a widespread phenomenon, which involves 30% of the genes we analyzed.

Recently there has been discussion on the role of cis-regulatory elements in the spatial organization of the genome and in their possible role in restricting chromosomal rearrangements (see [49] and review in [50]). The most well known example of this are the hox clusters, although they do exhibit wider plasticity in fish genomes than in other genomes. Our work shows clearly that shuffling of cis-regulatory elements is a widespread phenomenon within orthologous loci. It would be interesting to investigate further to what extent shuffling occurs on a genome-wide scale. Further analysis is required in order to understand the real extent of this phenomenon outside orthologous loci. This is the first genome-wide study in which we show that regulatory elements are mobile across species and that this needs to be taken in consideration when using comparative evolutionary methods for locating potential regulatory elements. It would be useful to assess the extent of shuffling genome-wide to develop a thresholding statistic. We have investigated this by searching for SCEs in fugu non-orthologous loci. Although

this results in a significantly lower number of hits (23100 hits in orthologous analysis, 9884 in nonorthologous analysis; p-value < 2.2e-16) , the result shows that shuffling does occur outside of the orthologous locus. It is difficult to interpret this result without taking into account other data (e.g. expression data and sequence similarity for genes considered as non orthologous or, indeed, in vivo assays on hits in non orthologous loci) that would allow us to establish to what extent hits in non orthologous loci are noise and to what extent they represent regulatory elements in genes with similar expression patterns. Finally we need to underline that the fact that our mammalian rCNE dataset is built using a global alignment approach will limit the search space and will not allow us to investigate the extent of regulatory element shuffling within mammals. This data reduction step has been used in the past [51], and it was used in our case based on the assumption that shuffling of regulatory elements is more likely to occurs over longer evolutionary distances. Widespread shuffling of elements could act as a potential mechanism for providing new expression sites to genes which are placed into the vicinity of a translocated enhancer. These issues can only be tackled appropriately by performing further analysis of the extent to which conserved elements shuffle beyond their locus of origin on both small and large evolutionary distances.

Conclusions

Our work shows that shuffling of cis-regulatory regions is a widespread phenomenon across the vertebrate lineage that affects approximately 70% of the conserved non-coding elements identified. The approach used allowed us to demonstrate that there is an order of magnitude more conserved elements in the vertebrate lineage than previously shown. Moreover conservation of regulatory elements occurs over thousands, rather than

hundreds of genes. By casting a wider net over vertebrate non-coding conservation we were able to demonstrate that there are hundreds of genes that do not belong to the "trans-dev" category, such as genes found in the membrane and extra-cellular regions, which also contain conserved non-coding elements. Finally, our in-vivo assays prove that although we cast a wider net, the catch was still as rich: more than 80% of the elements tested acted as enhancers, and the majority of them showed tissue-restricted patterns of expression in line with the neighboring gene.

Materials and Methods

Selection of genes and sequences

Groups of homologous genes from the genomes of *Mus musculus*, *Homo sapiens*, *Canis familiaris*, *Rattus norvegicus*, *Takifugu rubripes*, *Tetraodon nigroviridis* and *Danio rerio* were selected from the Ensembl-compara database [13] and their sequences were obtained from the Ensembl database release 32 [14]. Genes were considered homologous if they were classified as best reciprocal hits (BRH) in Ensembl-compara. We analyzed all the genes that were conserved in at least 4 species, of which 3 had to be human, mouse and fugu, and one could be either dog or rat. This selection led to 9,749 groups of homologous genes. For each gene we analyzed the whole genomic repeat-masked sequence containing the transcriptional unit as well as the complete flanking sequences up to the next gene upstream and the next gene downstream. The region was extracted from Ensembl and the 5' -> 3' sequence of the locus was stored in a custom database (i.e. all mouse genes were stored as being in forward strand on the sequences stored). In cases where the Ensembl gene contained multiple transcripts, the longest transcript was taken

in consideration for the pre, post and intron assignments of SCEs, but all exons (including those of other transcripts) were used to mask the sequence from coding regions. Similarly, if there were nested genes present in the locus, they were not taken in consideration to determine the extent of sequence to analyze, but they were taken in consideration to mask coding sequences in the region.

Identification of mammalian rCNEs

Global multiple alignments among human, mouse rat and dog were performed on each group of homologous genes using MLAGAN [25] with default parameters. The multiple alignments thus obtained were parsed using VISTA [31] with a window of 50 bases searching for conserved segments of at least 100 bp having a percentage identity of at least 70%. From these regions we selected as rCNEs only regions shared and overlapping in at least mouse, human and a third mammalian genome (either dog or rat) with a minimum length of 50 bp. In cases where the upstream region of an analyzed gene coincided with the downstream region of another analyzed gene rCNEs were counted only once.

Identification of SCEs

Mouse rCNEs were used as query sequences against the respective fugu, zebrafish and tetraodon homologous sequences using CHAOS [24] on both strands with the following parameters: word length=10, score cut-off=10, rescoring cut-off =1000, blast-like extension=on. Other parameters were left as set by default including the degeneracy tolerance of 1, i.e. allowing a single mismatch in the seed of the alignment. The hits thus obtained were filtered to retain only those with at least 60% identity and 40 bp length. Although three genomes were queried a hit in Fugu was required to consider the result an

SCE. All other hits (if any) were used to select the region of overlap as the final SCE, but only SCEs greater than 20bp after the overlap analysis were taken into consideration.

GO analysis

Ensembl gene IDs were converted into the corresponding RefSeq IDs before the analysis. The GOstat program [33] was used to find statistically over-represented GOs in the groups of genes, using the 'goa_mouse' GO gene association database as a comparator. The false discovery rate and the p-value cut-off of 0.001 options were used. Raw output was converted in supplementary tables using a custom Perl script. Simple association of genes to GO classes presented in figure 4 were produced using DAVID version 2 [52].

Mapping of conserved elements

rCNEs and SCEs were classified as "genic" if they overlapped any Ensembl genes , Ensembl EST genes [30], ESTs [53], EMBL proteins [54] or Genscan predictions [55] from the Ensembl *Mus musculus* genome build release 32. Furthermore each rCNE and SCE was classified with respect to the gene structure as "pre-gene", "intronic" and "post-gene" based on its location within these three portions of the locus. According to this "gene-centric" classification as well as the strand of the fugu CHAOS hits (since all genes were stored in forward strand) SCEs were classified as "collinear" (i.e. not changed in orientation and not shifted between gene portions), "moved" (shifted between gene portions), "reversed" (changed in orientation, but retained in the same gene portion) and "moved-reversed" (changed in orientation as well as shifted in gene portion).

Blast vs. CHAOS comparison

A subset of ~50% of the mammalian rCNEs were used as query sequences against the corresponding fugu homologous sequences using CHAOS [24] and Blast2 [16], using a gap penalty of 2 as was used in the CNE analysis and e-value set at infinity to ensure that

no hits would be filtered because of their statistical significance, analyzing both strands. The analysis was conducted three times varying only the word length used between 20, 15 and 10. The hits thus obtained were filtered in order to take only those sharing an identity of minimum 60% and a length of at least 40 bp.

Overlap analysis

Overlaps amongst different classes of conserved non-coding regions were defined using their genomic coordinates after having mapped all elements on the mouse loci used in this analysis. Since there is no downloadable dataset for CNGs, they were obtained by querying the GALA database [56] for conserved regions shared between human an mouse of at least 100bp and 70% identity. CNEs [15], UCEs [14] and known enhancers were downloaded from Genbank. Enhancers were downloaded by searching for enhancer features in mouse Genbank records and then checking them manually to eliminate mis-annotated entries. All the sequences thus downloaded were then mapped on the mouse loci used in our analysis by using Megablast [57] with default parameters for CNGs, UCEs and known enhancers and with a gap penalty of 2 for mapping CNEs, in accordance to the parameters used by Woolfe et al in their analysis. Elements were considered mapped with 75% coverage and 75% percentage identity. Only elements which did not map to exons were taken into consideration.

Identification of control fragments

A set of control fragments to be tested in vivo was built from the same gene loci in which the tested SCEs were found, by selecting regions which were not conserved and did not present repeats, of the same length and number as the elements tested.

Zebrafish embryo injections

The enhancer activity was assayed in conjunction with the minimal promoter *mHSP68*, which was previously shown to have low activity in zebrafish embryos and which has allowed the detection of enhancer function from several heterologous gene elements [28, 58]. HSP68lacZ-pBS DNA plasmids containing the mouse HSP68 promoter [58] and lacZ were prepared using the Promega PureYield Plasmid Midiprep System plasmid preparation kit, digested by Promega BamHI enzyme and DNA fragments were gel purified using the Promega Wizard SV Gel and PCR Clean-Up System kit. HSP-lacZ DNA fragments were re-suspended in 1% phenol red containing nuclease-free water at the concentration of 25 ng/µl as described[59] and were injected into the cytoplasm of zebrafish embryos at one cell stage. Wild type embryos (Tubingen AB) were collected after fertilization and dechorionated by pronase as described [60]. Fugu DNA was used for production of SCE fragments. Fragments were amplified by PCR, then isolated and purified using the Qiagen Qiex DNA purification kit and finally eluted in sterile water. For injection phenol red was added to have a final concentration of 50 ng/µl. Coinjection of PCR fragments at a concentration of 50 ng/µl reaching a range of 5 to 1 molar ratio with the HSP lacZ fragment. Embryos were maintained at 28° C and collected at prim 6 stage [61] and fixed and *lacZ* stained as described [27].

Analysis of transgene expression

LacZ stained embryos were analyzed by plotting the mosaic expression activity on expression maps as described [62, 63]. The co-injection experiments were repeated 3 times. Data form approximately 100–120 embryos were collected on a single expression map providing an expression profile. For each embryo expressing *lacZ* the number of expressing cells was counted and classified in muscle, notochord, CNS, eye, ear and

vessels. These tissues were selected because they are well defined at the time of inspection [27]. Other tissues which were either difficult to determine or might have represented abnormalities (ectopic tissue growth, apoptotic mismigrating cells) were counted as "other". 23 SCEs, 4 SCEs overlapping known mouse enhancers, 12 control fragments, 1 negative control consisting only of the *HSP:lacZ* fragment and a positive control, ArC [64], were analyzed. We verified the significance of the enhancement of expression over the general low level improvement of expression of co-injected fragments likely caused by carrier DNA effect (see e.g.) in two ways. Firstly, we aimed at detecting tissue-restricted enhancers, and secondly to identify generic enhancers. To identify tissue-restricted enhancers we compare, for each fragment co-injected, and for each tissue, the number of expressing cells with respect to the number of expressing cells from the embryos injected with the negative control in the respective tissues, only when the average of cells expressing *lacZ* in injected embryos was higher than in the control. Fisher exact tests were then used on the comparisons and a p-value cut-off of 0.01 was used to classify a fragment as a tissue-restricted enhancer. The identification of generic enhancers was performed by establishing the average and standard deviation of the number of expressing cells per expressing embryo in the control fragments and then classifying as enhancers fragments in which the number of expressing cells per embryo was higher than the average plus twice the standard deviation of the control fragments. In the calculation of the average and standard deviation we excluded the UBL7 control fragment, because it was a clear outlier which presented activity that was higher than any of the enhancers tested, including the positive control. All fragments classified as enhancers by either of the two tests were considered as positive.

Additional Data Files

The dataset described in this analysis is available at http://valis.tigem.it/sce.html for full download as well as search for SCEs belonging to individual genes.

Acknowledgements

We appreciate the useful input from two anonymous referees and we would like to acknowledge helpful discussions with Michael Brudno, Caterina Missero, Diego Di Bernardo, Marco Sardiello, Maria Luisa Chiusano, Giovanni Colonna and Roberto di Lauro. We would also like to thank for their technical support Marco De Simone, Mario Traditi and Alessandro Davassi. A special acknowledgement also goes to the late Parvesh Mahtani, who shared our enthusiasm for this project. This work was supported by the Fondazione Telethon and the Sixth Framework Program of the European Commission (LSH-2003-1.1.0-1).

References

1. Blackwood EM, Kadonaga JT: **Going the distance: a current view of enhancer action.** *Science* 1998, **281:**60-63.
2. Oda-Ishii I, Bertrand V, Matsuo I, Lemaire P, Saiga H: **Making very similar embryos with divergent genomes: conservation of regulatory mechanisms of Otx between the ascidians Halocynthia roretzi and Ciona intestinalis.** *Development* 2005, **132:**1663-1674.
3. Dickmeis T, Muller F: **The identification and functional characterisation of conserved regulatory elements in developmental genes.** *Brief Funct Genomic Proteomic* 2005, **3:**332-350.
4. Chuzhanova NA, Krawczak M, Nemytikova LA, Gusev VD, Cooper DN: **Promoter shuffling has occurred during the evolution of the vertebrate growth hormone gene.** *Gene* 2000, **254:**9-18.
5. Kermekchiev M, Pettersson M, Matthias P, Schaffner W: **Every enhancer works with every promoter for all the combinations tested: could new regulatory pathways evolve by enhancer shuffling?** *Gene Expr* 1991, **1:**71-81.
6. Surguchov A: **Migration of promoter elements between genes: a role in transcriptional regulation and evolution.** *Biomed Sci* 1991, **2:**22-28.
7. Boffelli D, Nobrega MA, Rubin EM: **Comparative genomics at the vertebrate extremes.** *Nat Rev Genet* 2004, **5:**456-465.
8. Dermitzakis ET, Reymond A, Antonarakis SE: **Conserved non-genic sequences - an unexpected feature of mammalian genomes.** *Nat Rev Genet* 2005, **6:**151-157.
9. Glazko GV, Koonin EV, Rogozin IB, Shabalina SA: **A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions.** *Trends Genet* 2003, **19:**119-124.
10. Sorek R, Ast G: **Intronic sequences flanking alternatively spliced exons are conserved between human and mouse.** *Genome Res* 2003, **13:**1631-1637.
11. Weber MJ: **New human and mouse microRNA genes found by homology search.** *Febs J* 2005, **272:**59-73.
12. Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, Krumlauf R, Brenner S: **Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes.** *Proc Natl Acad Sci U S A* 1995, **92:**1684-1688.
13. Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV, Antonarakis SE: **Numerous potentially functional but non-genic conserved sequences on human chromosome 21.** *Nature* 2002, **420:**578-582.
14. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome.** *Science* 2004, **304:**1321-1325.

15. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al: **Highly conserved non-coding sequences are associated with vertebrate development.** *PLoS Biol* 2005, **3:**e7.
16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
17. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci U S A* 1988, **85:**2444-2448.
18. Bergman CM, Kreitman M: **Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences.** *Genome Res* 2001, **11:**1335-1345.
19. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434:**338-345.
20. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al: **Aligning multiple genomic sequences with the threaded blockset aligner.** *Genome Res* 2004, **14:**708-715.
21. King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC: **Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences.** *Genome Res* 2005, **15:**1051-1060.
22. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res* 2005, **15:**1034-1050.
23. Ettwiller L, Paten B, Souren M, Loosli F, Wittbrodt J, Birney E: **The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates.** *Genome Biol* 2005, **6:**R104.
24. Brudno M, Chapman M, Gottgens B, Batzoglou S, Morgenstern B: **Fast and sensitive multiple alignment of large genomic sequences.** *BMC Bioinformatics* 2003, **4:**66.
25. Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S: **Glocal alignment: finding rearrangements during alignment.** *Bioinformatics* 2003, **19 Suppl 1:**i54-62.
26. Muller F, Blader P, Strahle U: **Search for enhancers: teleost models in comparative genomic and transgenic analysis of cis regulatory elements.** *Bioessays* 2002, **24:**564-572.
27. Muller F, Chang B, Albert S, Fischer N, Tora L, Strahle U: **Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord.** *Development* 1999, **126:**2103-2116.
28. Rastegar S, Albert S, Le Roux I, Fischer N, Blader P, Muller F, Strahle U: **A floor plate enhancer of the zebrafish netrin1 gene requires Cyclops (Nodal) signalling and the winged helix transcription factor FoxA2.** *Dev Biol* 2002, **252:**1-14.
29. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S: **LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA.** *Genome Res* 2003, **13:**721-731.

30. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, et al: **Ensembl 2006.** *Nucleic Acids Res* 2006, **34:**D556-561.
31. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I: **VISTA: visualizing global DNA sequence alignments of arbitrary length.** *Bioinformatics* 2000, **16:**1046-1047.
32. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
33. Beissbarth T, Speed TP: **GOstat: find statistically overrepresented Gene Ontologies within a group of genes.** *Bioinformatics* 2004, **20:**1464-1465.
34. Sprague J, Clements D, Conlin T, Edwards P, Frazer K, Schaper K, Segerdell E, Song P, Sprunger B, Westerfield M: **The Zebrafish Information Network (ZFIN): the zebrafish model organism database.** *Nucleic Acids Res* 2003, **31:**241-243.
35. **The Zebrafish Information Network** [http://zfin.org/]
36. Walton RZ, Bruce AE, Olivey HE, Najib K, Johnson V, Earley JU, Ho RK, Svensson EC: **Fog1 is required for cardiac looping in zebrafish.** *Dev Biol* 2006, **289:**482-493.
37. Kudoh T, Tsang M, Hukriede NA, Chen X, Dedekian M, Clarke CJ, Kiang A, Schultz S, Epstein JA, Toyama R, Dawid IB: **A gene expression screen in zebrafish embryogenesis.** *Genome Res* 2001, **11:**1979-1987.
38. Kudoh T, Dawid IB: **Zebrafish mab21l2 is specifically expressed in the presumptive eye and tectum from early somitogenesis onwards.** *Mech Dev* 2001, **109:**95-98.
39. Zecchin E, Conigliaro A, Tiso N, Argenton F, Bortolussi M: **Expression analysis of jagged genes in zebrafish embryos.** *Dev Dyn* 2005, **233:**638-645.
40. Smale ST, Kadonaga JT: **The RNA polymerase II core promoter.** *Annu Rev Biochem* 2003, **72:**449-479.
41. Ludwig MZ, Bergman C, Patel NH, Kreitman M: **Evidence for stabilizing selection in a eukaryotic enhancer element.** *Nature* 2000, **403:**564-567.
42. Tautz D: **Evolution of transcriptional regulation.** *Curr Opin Genet Dev* 2000, **10:**575-579.
43. Poulin F, Nobrega MA, Plajzer-Frick I, Holt A, Afzal V, Rubin EM, Pennacchio LA: **In vivo characterization of a vertebrate ultraconserved enhancer.** *Genomics* 2005, **85:**774-781.
44. Adams MD: **Conserved sequences and the evolution of gene regulatory signals.** *Curr Opin Genet Dev* 2005, **15:**628-633.
45. Nobrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM: **Megabase deletions of gene deserts result in viable mice.** *Nature* 2004, **431:**988-993.
46. **Enhancer Browser** [http://enhancer.lbl.gov/]
47. Miles CG, Rankin L, Smith SI, Niksic M, Elgar G, Hastie ND: **Faithful expression of a tagged Fugu WT1 protein from a genomic transgene in zebrafish: efficient splicing of pufferfish genes in zebrafish but not mice.** *Nucleic Acids Res* 2003, **31:**2795-2802.

48. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, et al: **Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes.** *Science* 2002, **297:**1301-1310.
49. Liu Z, Garrard WT: **Long-range interactions between three transcriptional enhancers, active Vkappa gene promoters, and a 3' boundary sequence spanning 46 kilobases.** *Mol Cell Biol* 2005, **25:**3220-3231.
50. Pederson T: **The spatial organization of the genome in mammalian cells.** *Curr Opin Genet Dev* 2004, **14:**203-209.
51. Van Hellemont R, Monsieurs P, Thijs G, de Moor B, Van de Peer Y, Marchal K: **A novel approach to identifying regulatory motifs in distantly related genomes.** *Genome Biol* 2005, **6:**R113.
52. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4:**P3.
53. Boguski MS, Lowe TM, Tolstoshev CM: **dbEST--database for "expressed sequence tags".** *Nat Genet* 1993, **4:**332-333.
54. Cochrane G, Aldebert P, Althorpe N, Andersson M, Baker W, Baldwin A, Bates K, Bhattacharyya S, Browne P, van den Broek A, et al: **EMBL Nucleotide Sequence Database: developments in 2005.** *Nucleic Acids Res* 2006, **34:**D10-15.
55. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268:**78-94.
56. Giardine B, Elnitski L, Riemer C, Makalowska I, Schwartz S, Miller W, Hardison RC: **GALA, a database for genomic sequence alignments and annotations.** *Genome Res* 2003, **13:**732-741.
57. Zhang Z, Schwartz S, Wagner L, Miller W: **A greedy algorithm for aligning DNA sequences.** *J Comput Biol* 2000, **7:**203-214.
58. Kothary R, Clapoff S, Darling S, Perry MD, Moran LA, Rossant J: **Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice.** *Development* 1989, **105:**707-714.
59. Muller F, Lakatos L, Dantonel J, Strahle U, Tora L: **TBP is not universally required for zygotic RNA polymerase II transcription in zebrafish.** *Curr Biol* 2001, **11:**282-287.
60. Akimenko MA, Johnson SL, Westerfield M, Ekker M: **Differential induction of four msx homeobox genes during fin development and regeneration in zebrafish.** *Development* 1995, **121:**347-357.
61. Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF: **Stages of embryonic development of the zebrafish.** *Dev Dyn* 1995, **203:**253-310.
62. Müller F, Williams DW, Kobolak J, Gauvry L, Goldspink G, Orban L, Maclean N: **Activator effect of coinjected enhancers on the muscle-specific expression of promoters in zebrafish embryos.** *Mol Reprod Dev* 1997, **47:**404-412.
63. Müller F, Chang B, Albert S, Fischer N, Tora L, Strahle U: **Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord.** *Development* 1999, **126:**2103-2116.

64. Parks RJ, Bramson JL, Wan Y, Addison CL, Graham FL: **Effects of stuffer DNA on transgene expression from helper-dependent adenovirus vectors.** *J Virol* 1999, **73:**8027-8034.

Figure Legends

Figure 1 - Number of conserved gene loci vs number of rCNEs identified in the mouse, rat, human and dog genomes

Graph showing the number of rCNEs found conserved in the dog, rat, mouse and human genomes vs the number of genes found conserved across the same genomes. Though almost 90% of the genes can be found in all 4 genomes, most rCNEs can be found only in 3 out 4 genomes.

Figure 2 - Distribution of length, percentage identity and shuffling categories of SCEs

A/B/C/D: SCEs were categorized based on their change in location and orientation in Fugu rubripes with respect to their location and orientation in the mouse locus. The entire locus, comprising the entire flanking sequence up to the next up- and down-stream gene was taken in consideration. Definition of specific classes:

A. Collinear SCEs- elements that have not undergone any change in location or orientation within the entire gene locus

B. Reversed SCEs- elements that have changed their orientation in the fish locus with respect to the mouse locus, but have remained in the same portion of the locus.

C. Moved SCEs- elements that have moved between the pre-gene, post-gene and intronic portions of the locus.

D. Moved-reversed - elements that have undergone both of the above changes.

E. Frequency distribution of SCE length in base pairs.

F. Frequency distribution of percentage identity of SCE hits in fugu.

Figure 3 - Examples of loci containing shuffled conserved elements

A. the Sema6d (sema domain, transmembrane domain, and cytoplasmic domain, semaphorin 6D - MGI:2387661) locus contains a post-genic moved-reversed conserved element. The SCE is found downstream to the gene in mammalian loci and upstream of the gene in fish genomes, and in reverse orientation only in the genomes of fugu and tetraodon.

B. the Ptprg (protein tyrosine phosphatase, receptor type, G - MGI:97814) locus contains an intronic moved-reversed conserved element. The SCE is found in the first intron of the Ptprg gene in mammalian genomes, downstream of the gene in reverse orientation in fugu and tetraodon, and in the second intron in reverse orientation in zebrafish. Boxes represent the multiple alignments of the SCEs identified.

Figure 4 - GO Classification of genes harbouring CNEs vs. genes harbouring SCEs

All genes containing CNEs and/or SCEs were analyzed for GO (gene ontology) term classification. Genes containg CNEs are shown in red and genes containing SCEs are shown in gray. Plots show differences in absolute numbers as well as relative percentages. Classification is shown for cellular component (A) and biological process (B) categories.

Figure 5 - Analysis of SCE shuffling in 1000bp windows

Each column in the figure shows the analysis of a locus portion (pre-gene, intron-start, intron-end and post-gene) divided in 1000bp windows. In each column the first graph indicates the number of collinear SCEs identified, the second graph the number of non-collinear SCEs identified and the third graph the chi-square test used to identify windows which show a significant deviation from the expected proportion of collinear to non-collinear SCEs. The p-value is shown for the only window (1000bp upstream of TSS) which shows a significant deviation from the expected proportion.

Figure 6 - Overlap of known mouse enhancers with conserved elements

All mouse enhancers deposited in GenBank (94) were mapped to the genome and compared to previously published conserved elements (UCEs and CNEs) as well as our own dataset of SCEs to verify their overlap. Only 1 known mouse enhancer is overlapped by a CNE, and 2 by a UCE while our dataset of SCEs identifies 18 known mouse enhancers as being conserved within fish genomes

Figure 7 - Expression profiles of X-Gal stained embryos

A/B/C/D/E/F. Expression profiles of 1 day old X-Gal stained zebrafish embryos. Each expression map represents a composite overview of the LacZ positive cells of 65-175 embryos. Gene names and fragment/SCE id are shown. Detailed distribution of X-Gal stained cells in different tissues as well as data for all other fragments is shown in Table 3. Side view of head region of LacZ stained embryos are shown with anterior to the left. A. HSP-lacZ injected embryo D. Embryo co-injected with SCE 3121 associated with Jag1b gene. F. Embryo co-injected with SCE 4939 associated with Mab21l2 gene.

Tables

rCNE type[a]	Total[b]	Coding[c]	Non-coding[d]
Total[e]	364,358	82,714	281,644
pre_gene[f]	120,001	23,832	96,169
intronic[g]	158,722	29,002	129,720
post_gene[h]	85,521	29,766	55,755

Table 1 - Transcription potential, localization and number of mammalian regionally conserved non-coding elements (rCNE)

[a] Type of conserved non-coding sequence (rCNE).

[b] Total number of rCNEs comprising genic and non-genic (see c, d).

[c] Number of genic rCNEs: overlapping EMBL proteins, ESTs, GenScan predictions and Ensembl genes

[d] Number of non-genic rCNEs: not overlapping EMBL proteins, ESTs, GenScan and Ensembl genes

[e] Total number of rCNEs comprising pre_gene, intronic and post_gene (see f, g,h)

[f] Number of pre_gene rCNEs: rCNEs localized before the translation start of the reference gene.

[g] Number of intronic rCNEs: rCNEs localized within the introns of the reference gene.

[h] Number of post_gene rCNEs: rCNEs localized after the translation end of the reference gene.

SCE type[a]	Total[b]	Coding[c]	Non-coding[d]
Total[e]	27,196	5,769	21,427
pre_gene[f]	8,387	1,363	7,024
Intron[g]	11,657	1,838	9,819
post_gene[h]	7,152	2,568	4,584

Table 2 - Transcription potential, localization and number of vertebrate shuffled conserved elements (SCEs)

[a] Type of shuffled conserved sequence (SCE).

[b] Total number of SCEs comprising genic and non-genic (see c, d).

[c] Number of genic SCEs: overlapping EMBL proteins, ESTs, GenScan predictions and Ensembl genes

[d] Number of non-genic SCEs: not overlapping EMBL proteins, ESTs, GenScan and Ensembl genes

[e] Total number of SCEs comprising pre_gene, intronic and post_gene (see f, g,h)

[f] Number of pre_gene SCEs: SCEs localized before the translation start of the reference gene.

[g] Number of intronic SCEs: SCEs localized within the introns of the reference gene.

[h] Number of post_gene SCEs: SCEs localized after the translation end of the reference gene

Gene	trans dev	name	SCE bp	SCE class	ENH	embryo	cell	ce/ emb	muscle	notoc.	CNS	eye	ear	vessels	other
no	NA	lacZ			neg c	161	40	0.25	p-value	p-value	p-value	p-value	p-value	p-value	p-value
Shh	Y	ArC			pos c	96	242	**2.52**		**8.48E-07**					
Shh	Y	12058	45	rev	Y	139	69	0.5	**6.86E-09**						
Otx2	Y	13988	51	mov	Y	111	93	0.84	0.6444		**0.006269**	0.5536	0.3155		
Gata3	Y	15402	40	mre	Y	107	103	**0.96**			0.398	0.5764	0.1906		1
Ets	Y	8744	40	mov	Y	105	180	**1.57**			**0.002593**			**4.78E-09**	
Ets	Y	8745	46	mov	Y	133	210	**1.58**			0.1558	0.6015	0.3619	**2.15E-06**	
Ets	Y	8726	41	mre	Y	159	345	**2.17**			0.05534	0.6136	0.1485	**2.08E-06**	
Ets	Y	8728	48	mre	Y	149	176	**1.18**			0.0444	0.129	0.07924	**1.31E-05**	
Pax2b	Y	31027	39	col	Y	149	105	0.7			**0.002374**	0.06327	0.1902		
Pax6a	Y	15696	33	mov	Y	133	122	**0.92**			**8.21E-06**	0.3343	0.01268		
Pax3	Y	24781	42	mov	N	124	67	0.54	0.02982		0.5287	1			
Zfpm2	Y	23818	48	col	Y	140	119	0.85			**1.49E-06**	0.01296	1		
Zfpm2	Y	23838	48	mre	Y	131	148	**0.98**			**0.0003576**	0.04369	0.1231		
Tmeff2	N	26014	48	mov	N	164	125	0.76			0.7654	0.02301	0.3371		0.2801
Tmeff2	N	26015	38	mov	Y	120	159	**1.33**	**0.001035**		0.303	0.2088			
Tmeff2	N	26016	51	mre	Y	109	148	**1.36**			**0.0006309**	0.0149	0.5862		
Jag1b	Y	16407	37	col	N	136	98	0.72	1		0.1849	1	1		
Jag1b	Y	16408	55	col	Y	142	109	0.86			**5.45E-08**	**0.006524**	0.3245		
Jag1b	Y	16409	44	rev	N	106	54	0.51	1		0.5088	1	0.5058		
Mapkap1	N	17058	37	mov	Y	143	295	**2.06**	0.6825		0.05292	0.3788	0.6065		1
Mapkap1	N	17059	39	mov	Y	136	171	**1.26**	0.6686		**0.004037**	0.5973	0.077	0.5197	
Mab21l2	Y	23001	42	col	Y	142	317	**2.23**			**1.24E-07**	**0.004985**	0.2339		
Mab21l2	Y	23002	37	mre	Y	155	122	0.79			**7.85E-08**	**0.004138**			
Hmx3	Y	11669	150	col	Y	165	136	0.82			**0.001029**	0.07062	0.01423		
Lmx1b	Y	17027	300	col	Y	116	105	0.91			**0.00762**	0.1876	1		
3110004L20Rik	N	5803	45	mre	N	65	16	0.25	0.2929						1
3110004L20Rik	N	5802	39	mov	Y	122	320	**2.62**	0.1874	0.01209					
Elmo1	N	6026	45	Rev	Y	103	76	0.74	**0.007132**	0.6848					
Ets	Y	11216	NA	Ctrl	N	104	74	0.71	1						0.6954
Gata3	Y	3255	NA	Ctrl	N	174	110	0.63	0.04481		0.281	0.5739	0.02163		
1300007F04Rik	N	2797	NA	Ctrl	N	157	115	0.73							
Tmeff2	N	198	NA	Ctrl	N	145	23	0.16	0.7448		0.6597		0.3651		
Mab21l2	Y	909	NA	Ctrl	N	165	92	0.56	0.06359		1	1	1		
3110004L20Rik	N	410	NA	Ctrl	N	107	23	0.21							0.01984
Elmo1	N	10157	NA	Ctrl	N	146	38	0.26	0.287	0.8126					
Shh	Y	11271	NA	Ctrl	Y	165	83	0.5	**3.34E-07**		1	1	1		
Impact	Y	5990	NA	Ctrl	N	150	101	0.67	0.6496		0.2754		0.0622		
Ubl7	N	268	NA	Ctrl	Y	117	644	**5.5**	**0.0003325**		**7.15E-11**	0.02555	0.6197		
Lmx1b	Y	11767	NA	Ctrl	N	116	15	0.13	0.2743				0.0707		1
Irx3	Y	5945	NA	Ctrl	N	93	15	0.16	0.03938						

Table 3 - Analysis of X-Gal staining in zebrafish embryos co-injected with the Hsp promoter and SCEs or control fragments

For each DNA fragment tested the following information is given, from left to right: the gene locus in which the DNA fragment is found, indication about the GO classification of the gene in the 'trans-dev' class (Y = yes, N = no), the identifier given to the SCE or control fragment, the size of the SCE, the class (rev = reversed, mov = moved, mre = moved and reversed, col = collinear, Ctrl = control), summary about the potentially enhancer function of the element (Y = yes, N = no), the number of embryos injected, the total number of cells X-gal-stained, the ratio of stained cells divided by the number of embryos observed (bold and underlined those showing significant generalized enhancer activity), the p-value indicating the significance of the number of cells observed in the fragment tested versus the *lacZ:HSP* control for each tissue (bold and underlind for p-values < 0.01, see Materials and Methods). See Supplementary Table S3 for further info on fragments tested.

Legends for Additional Files

Figure S1 - Boxplots comparing the distribution of the distance of collinear vs. non-collinear NON-GENIC SCEs from the transcriptional unit

The four boxplots (from left to right) represent the distance of: pre-gene elements from the TSS, intronic SCEs from the start of the intron, post-gene SCEs from the the 3' end of the transcriptional unit, intronic elements from the end of the intron. The p-value obtained by performing a Wilcoxon rank test for the difference between the collinear and linear distributions is shown in each sub-figure.

Figure S2 - Venn diagram showing the overlap analysis of 4 datasets: CNGs, UCEs, CNEs and SCEs

Overlap analysis comparing our dataset of SCEs (shuffled conserved elements) with previously published datasets of conserved elements, as described below (ordered by dataset size). UCEs: ultra-conserved elements, identified using BLAST by selecting regions conserved between human and mouse, of minimum length 200bp and 100% identity [17]; CNEs: conserved non-coding elements, identified using MEGABLAST (with word size 20) by selecting all regions found conserved between fugu and human longer than 100bp. [12]; SCEs: shuffled conserved elements identified in our study using a global-local strategy combining MLAGAN alignment of mammalian loci and CHAOS alignment of mammalian conserved regions against the orthologous fish loci (see methods for details), with minimum length 40bp, and minimum identity 60%; CNGs: conserved non-genic elements, identified using BLASTZ by selecting regions conserved between human and mouse of minimum length 100bp and minimum 70% identity. [16]

Figure S3 - Graph showing the number and type of conserved elements identified by CHAOS and BLAST2 in our dataset as a function of the word size used

Graphs showing the number of conserved elements identified by CHAOS and BLAST2 in our dataset as a function of the word size used (at word sizes 5,10,20). All elements identified were filtered for minimum 40bp length and 60% identity. CHAOS scales exponentially in the number of elements identified as word size is diminished, resulting in an order of magnitude difference in the number of hits found for each decrease of 5 letters in word size. On the other hand results obtained with BLAST2 remain almost unaltered. This is particularly true of elements that have undergone shuffling, as shown by labeling elements with different colors based on the "shuffling class": Black = collinear (no shift in position or orientation between mouse and fugu); Light Blue = moved (shifted in position but not orientation); Pink = reversed (shifted in orientation); Dark Blue = moved_reversed (shifted in orientation and and position)

Table S1 - GO analysis of genes associated with CNEs and genes associated with SCEs

GOStat analysis to detect significant over- and under-representation ($p < 0.001$) of genes containing SCEs, genes containing CNEs in comparison with each other as well as in comparison with the dataset of all genes analyzed

Table S2 - GO analysis results for genes associated with collinear non-genic SCEs located 1000 bp upstream of the TSS

GOstat comparison of genes containing collinear non-genic SCEs with all analyzed genes.

Table S3 - Further information on all fragments tested

Table providing detailed information on fragments tested: gene name, fragment id, ensembl identifier of the gene, size of the pcr product injected, size of the SCE, location in the mouse genome, location and orientation in the fugu genome, and number of positively X-Gal stained cells per tissue analyzed.

File S1 - Supplementary information about tested fragments containing SCE

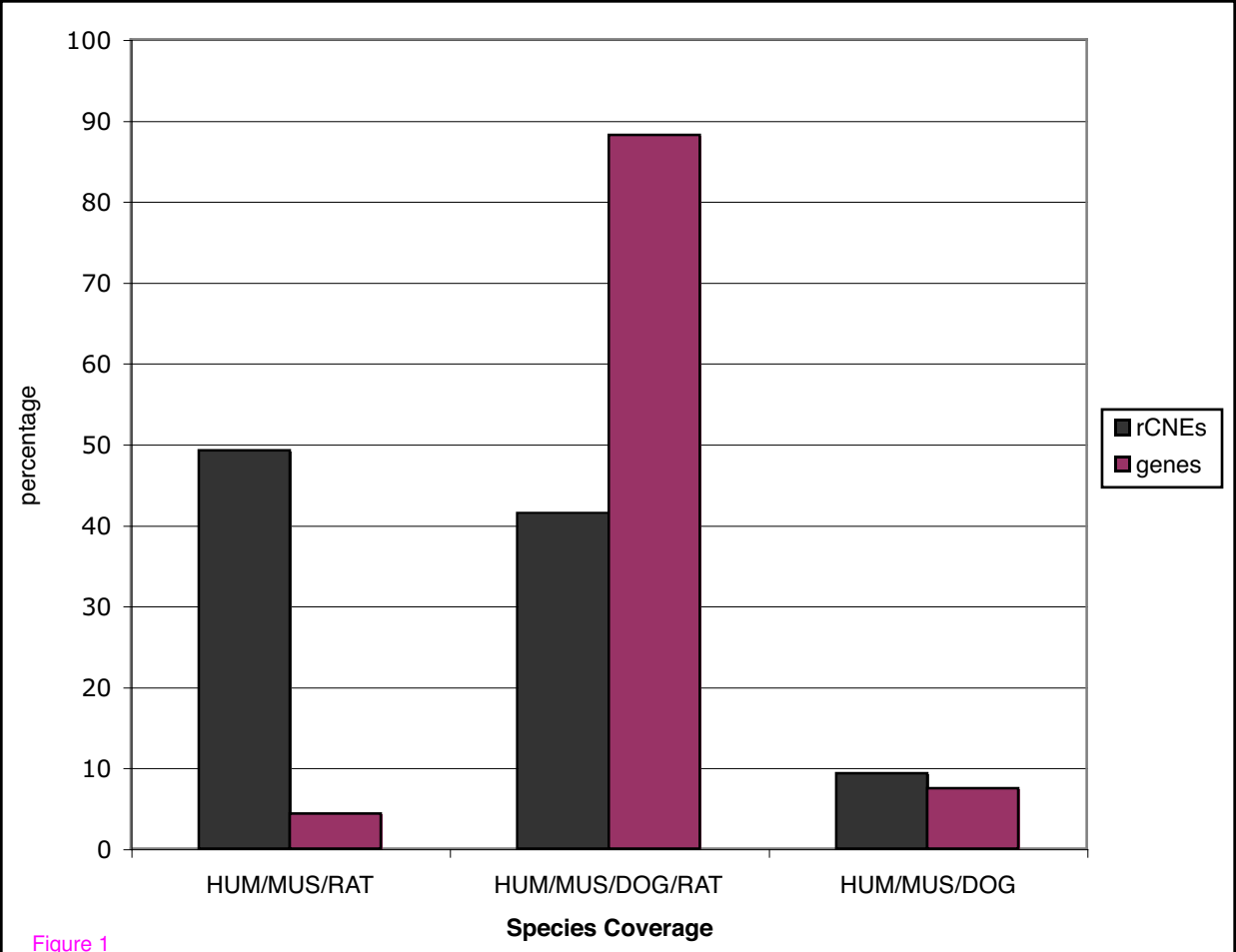
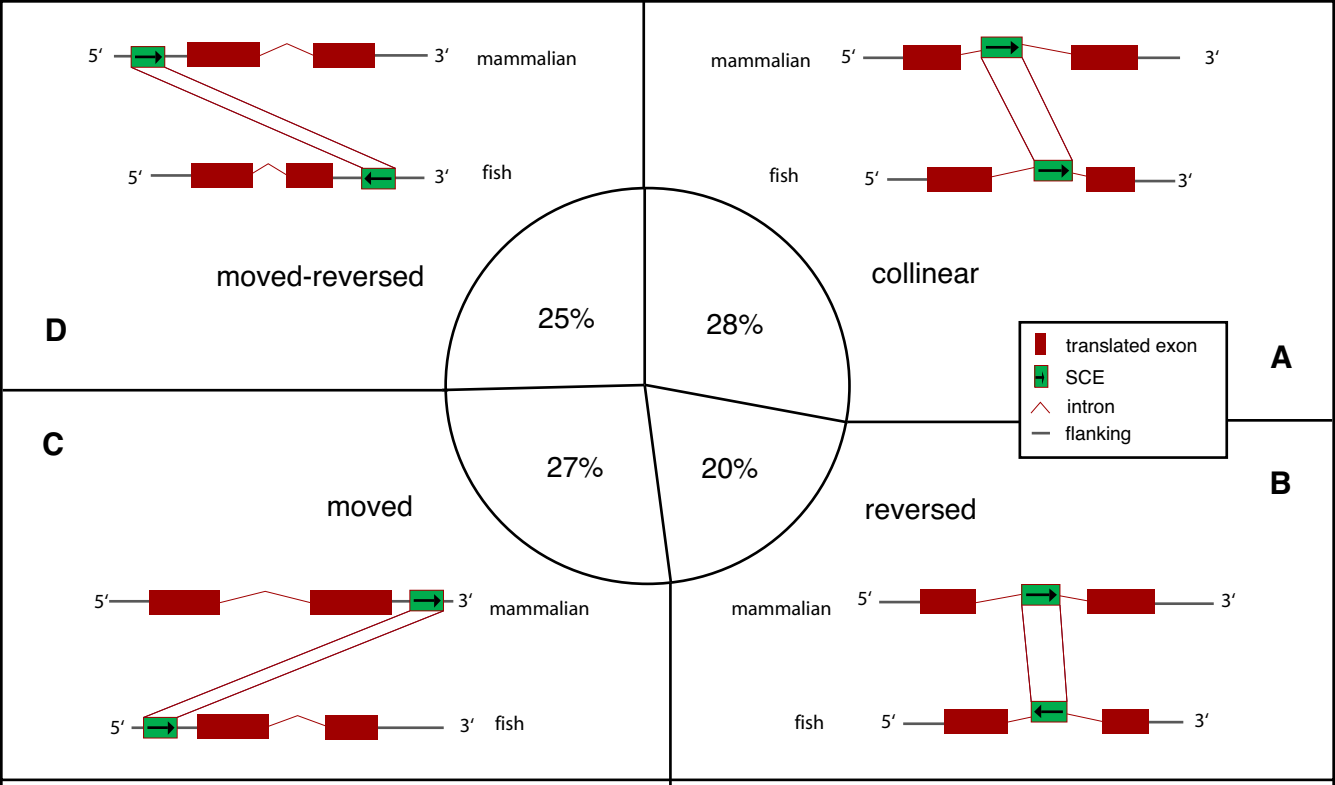
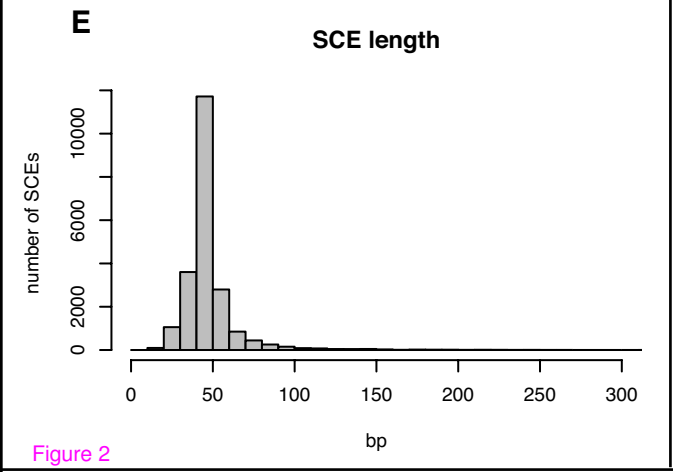

100
90
80
70
60
50
40
30
20
10
0
percentage
rCNEs
genes
HUM/MUS/RAT
HUM/MUS/DOG/RAT
HUM/MUS/DOG
Species Coverage

Figure 1


5′
3′
mammalian
5′
3′
fish
moved-reversed
D
25%
collinear
28%
mammalian
5′
3′
fish
5′
3′
translated exon
SCE
intron
flanking
A
C
27%
20%
B
moved
reversed
5′
3′
mammalian
5′
3′
fish
mammalian
5′
3′
fish
5′
3′


E
SCE length
number of SCEs
0
2000
6000
10000
0
50
100
150
200
250
300
bp

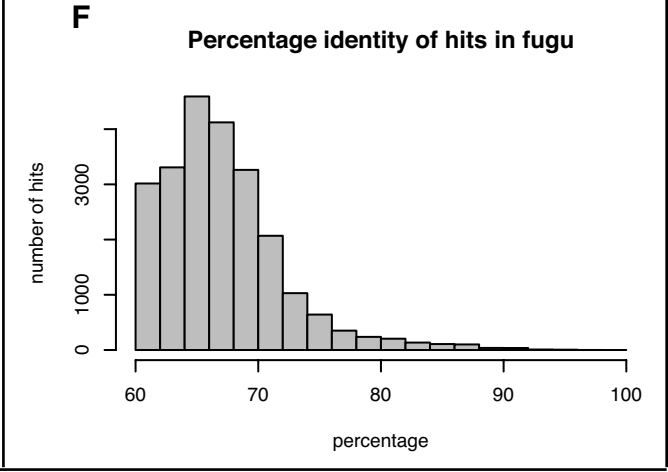

F
Percentage identity of hits in fugu
number of hits
0
1000
3000
60
70
80
90
100
percentage

Figure 2

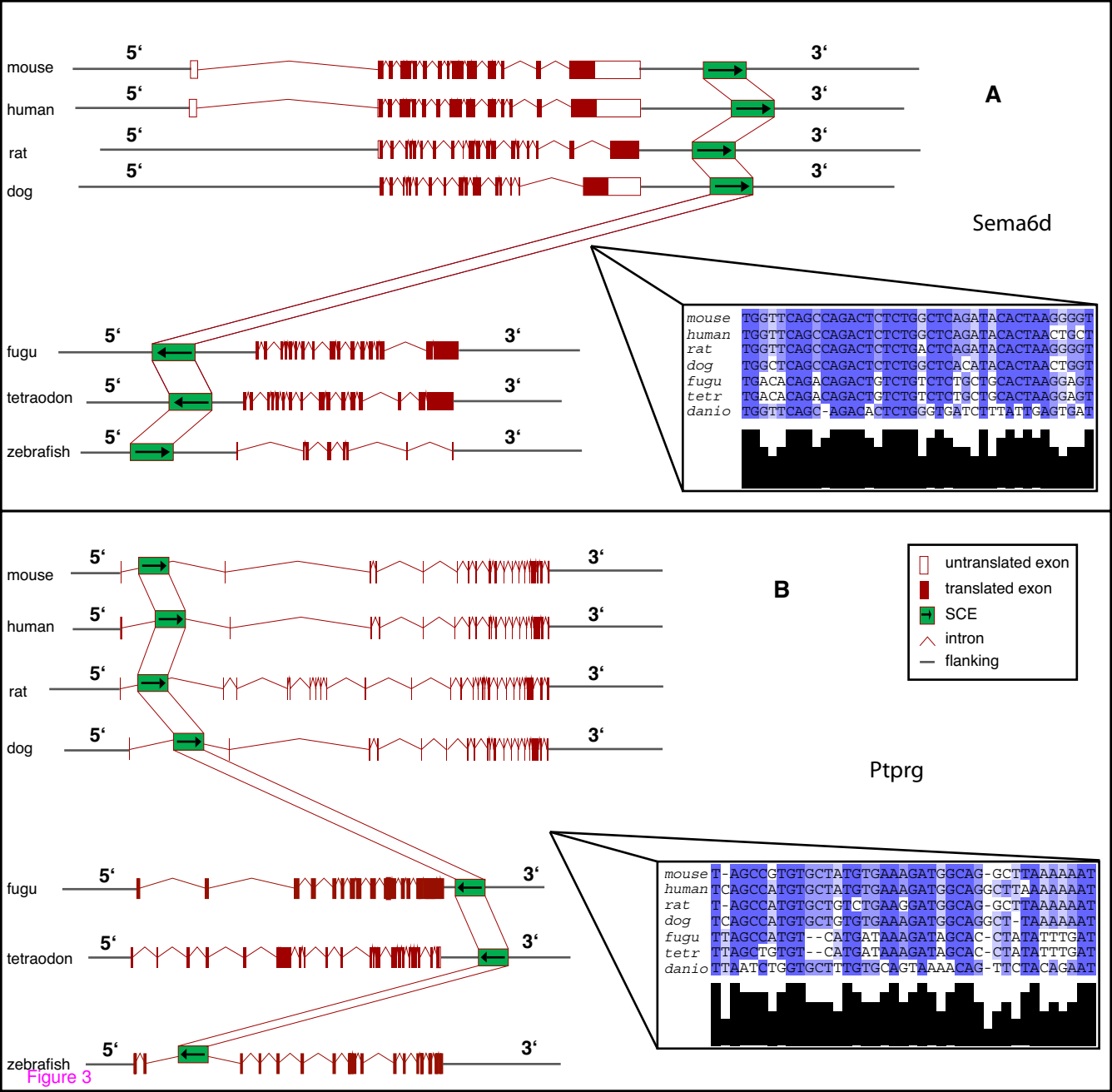

A
mouse
human
rat
dog
fugu
tetraodon
zebrafish
5‘
3‘
Sema6d
mouse TGGTTCAGCCAGACTCTCTGGCTCAGATACACTAAGGGGT
human TGGTTCAGCCAGACTCTCTGGCTCAGATACACTAACTGCT
rat TGGTTCAGCCAGACTCTCTGACTCAGATACACTAAGGGGT
dog TGGCTCAGCCAGACTCTCTGGCTCACATACACTAACTGGT
fugu TGACACAGACAGACTGTCTGTCTCTGCTGCACTAAGGAGT
tetr TGACACAGACAGACTGTCTGTCTCTGCTGCACTAAGGAGT
danio TGGTTCAGC-AGACACTCTGGGTGATCTTTATTGAGTGAT
B
untranslated exon
translated exon
SCE
intron
flanking
Ptprg
mouse T-AGCCGTGTGCTATGTGAAAGATGGCAG-GCTTAAAAAAT
human TCAGCCATGTGCTATGTGAAAGATGGCAGGCTTAAAAAAAT
rat T-AGCCATGTGCTGTCTGAAGGATGGCAG-GCTTAAAAAAT
dog TCAGCCATGTGCTGTGTGAAAGATGGCAGGCT-TAAAAAAT
fugu TTAGCCATGT--CATGATAAAGATAGCAC-CTATATTTGAT
tetr TTAGCTGTGT--CATGATAAAGATAGCAC-CTATATTTGAT
danio TTAATCTGGTGCTTTGTGCAGTAAAACAG-TTCTACAGAAT

Figure 3

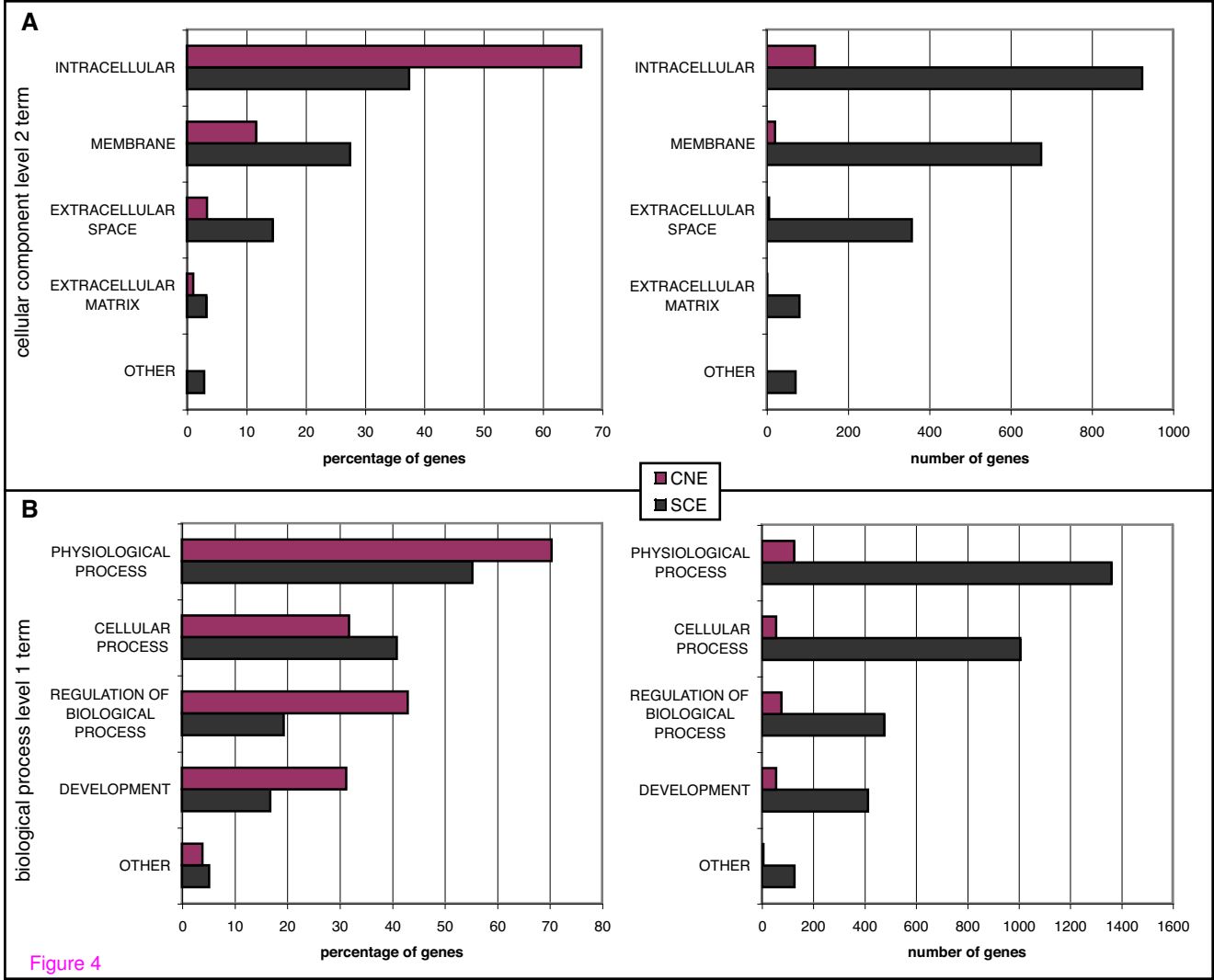

A
cellular component level 2 term
INTRACELLULAR
MEMBRANE
EXTRACELLULAR SPACE
EXTRACELLULAR MATRIX
OTHER
0 10 20 30 40 50 60 70
percentage of genes
INTRACELLULAR
MEMBRANE
EXTRACELLULAR SPACE
EXTRACELLULAR MATRIX
OTHER
0 200 400 600 800 1000
number of genes
CNE
SCE
B
biological process level 1 term
PHYSIOLOGICAL PROCESS
CELLULAR PROCESS
REGULATION OF BIOLOGICAL PROCESS
DEVELOPMENT
OTHER
0 10 20 30 40 50 60 70 80
percentage of genes
PHYSIOLOGICAL PROCESS
CELLULAR PROCESS
REGULATION OF BIOLOGICAL PROCESS
DEVELOPMENT
OTHER
0 200 400 600 800 1000 1200 1400 1600
number of genes

Figure 4

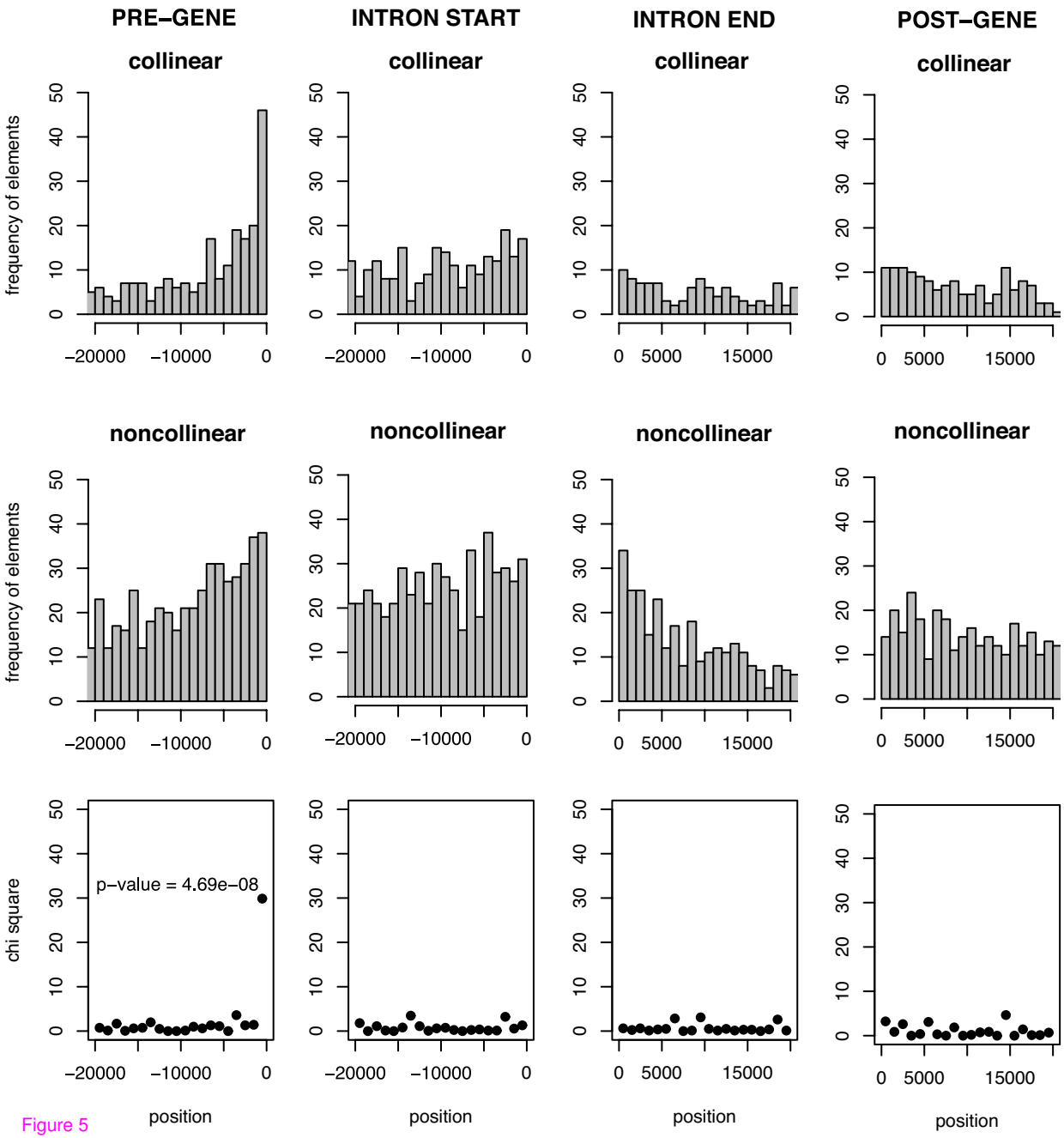

PRE–GENE
INTRON START
INTRON END
POST–GENE
collinear
noncollinear
frequency of elements
chi square
p–value = 4.69e–08
position
–20000
–10000
0
5000
15000

Figure 5

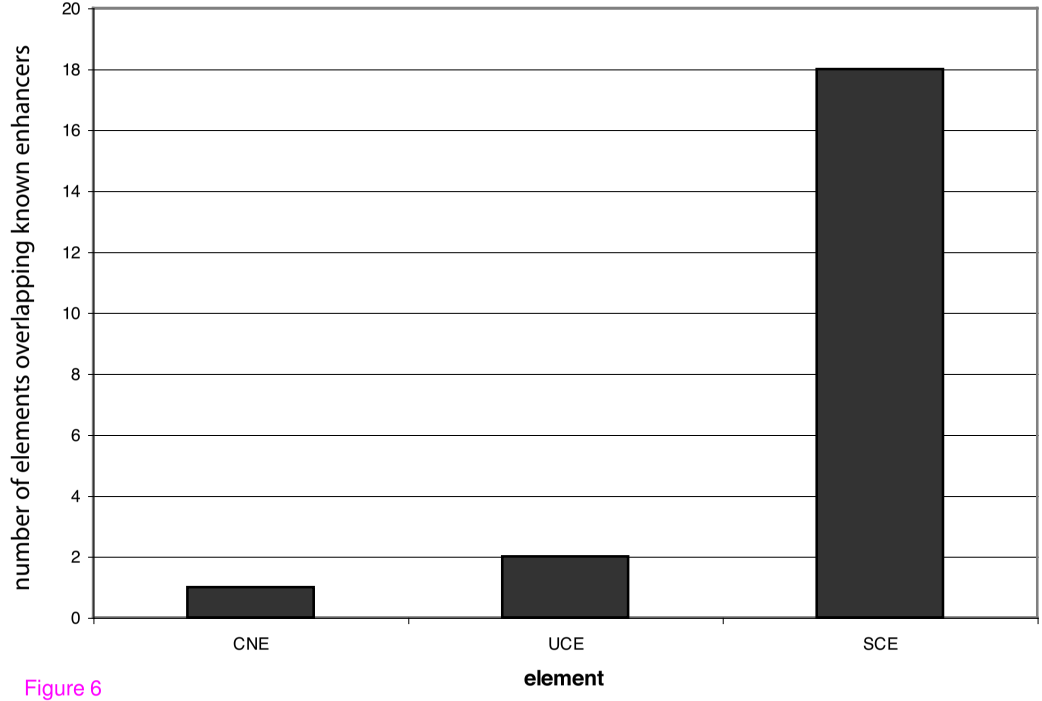

number of elements overlapping known enhancers
20
18
16
14
12
10
8
6
4
2
0
CNE
UCE
SCE
element

Figure 6

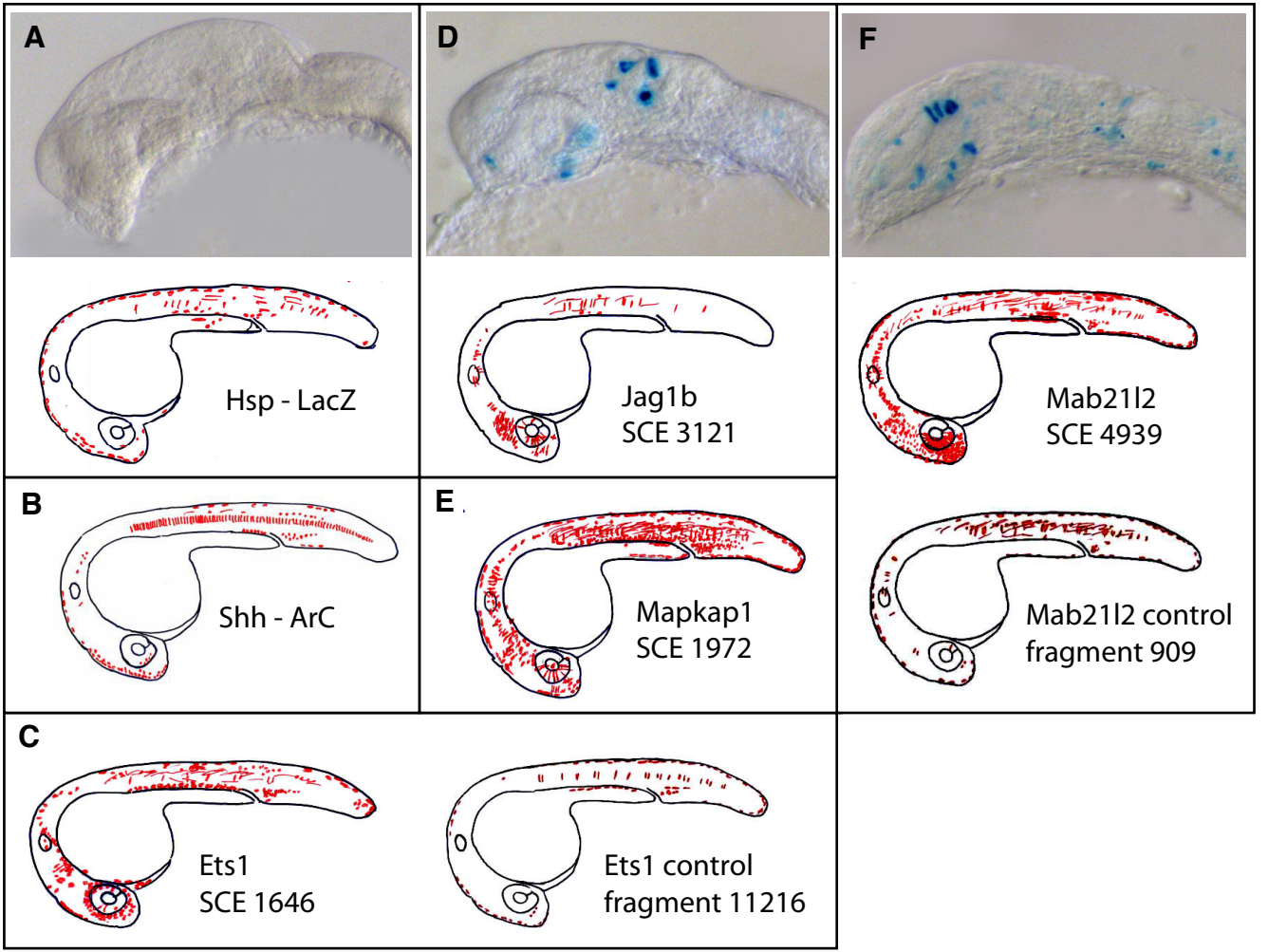

A
D
F
Hsp - LacZ
Jag1b
SCE 3121
Mab21l2
SCE 4939
B
E
Shh - ArC
Mapkap1
SCE 1972
Mab21l2 control
fragment 909
C
Ets1
SCE 1646
Ets1 control
fragment 11216

Figure 7

Additional files provided with this submission:

Additional file 7 : FileS1.pdf : 470Kb
http://genomebiology.com/imedia/5576616779877802/sup7.PDF
Additional file 6 : tabS3.xls : 22Kb
http://genomebiology.com/imedia/1595117090102288/sup6.XLS
Additional file 5 : TableS2.xls : 17Kb
http://genomebiology.com/imedia/7832912889877802/sup5.XLS
Additional file 4 : TableS1.xls : 32Kb
http://genomebiology.com/imedia/6765310269877802/sup4.XLS
Additional file 3 : FigureS3.eps : 644Kb
http://genomebiology.com/imedia/1428271418102280/sup3.EPS
Additional file 2 : FigureS2.eps : 2248Kb
http://genomebiology.com/imedia/4034773511022808/sup2.EPS
Additional file 1 : FigureS1.eps : 813Kb
http://genomebiology.com/imedia/8681723301022808/sup1.EPS