In Vivo Enhancer Analysis of Human Conserved Noncoding Sequences

Len A. Pennacchio[1,2,*], Nadav Ahituv[1,2], Alan M. Moses[2], Shyam Prabhakar[2], Marcelo Nobrega[2], Malak Shoukry[2], Simon Minovitsky[2], Inna Dubchak[1,2], Amy Holt[2], Keith D. Lewis[2], Ingrid Plajzer-Frick[2], Jennifer Akiyama[2], Sarah De Val[4], Veena Afzal[2], Brian L. Black[4], Olivier Couronne[1,2], Michael B. Eisen[2,3], Axel Visel[2], and Edward M. Rubin[1,2]

[1]U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598 USA.

[2]Genomics Division, MS 84-171, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 USA.

[3]Molecular and Cellular Biology Department, University of California-Berkeley, CA. 954720. USA.

[4]Cardiovascular Research Institute, University of California, San Francisco, CA 94143-2240, USA.

*To whom correspondence should be addressed: Len A. Pennacchio, Genomics Division, One Cyclotron Road, MS 84-171, Lawrence Berkeley National Laboratory, Berkeley, CA 94720. Email: LAPennacchio@lbl.gov, Phone: (510) 486-7498, Fax: (510) 486-4229.

Abstract

The identification of enhancers with predicted specificities in vertebrate genomes remains hampered by a lack of experimentally validated training sets. In this study, we leveraged extreme evolutionary sequence conservation as a filter to identify putative gene regulatory elements and characterized the *in vivo* enhancer activity of a large group of human-fish conserved and ultraconserved[1] noncoding elements in the human genome. We tested 167 of these extremely conserved sequences in a transgenic mouse enhancer assay and observed that 45% functioned reproducibly as tissue-specific enhancers of gene expression at embryonic day 11.5. While directing expression in a broad range of anatomical structures in the embryo, the majority of the 75 enhancers directed expression to various regions of the developing nervous system. We identified sequence signatures enriched in a subset of these elements that targeted forebrain expression and used these features to rank all ~3,100 human-fugu conserved noncoding elements in the human genome. The testing of the top predictions in transgenic mice resulted in a three-fold enrichment for sequences with forebrain enhancer activity. These data dramatically expand the catalogue of *in vivo*-characterized human gene enhancers and illustrate the utility of such training sets for a variety of biological applications including decoding the regulatory vocabulary of the human genome.

Identifying the sequences that direct the spatial and temporal expression of genes and defining their function *in vivo* is a major challenge in the annotation of vertebrate genomes. Significant progress has been made in the identification of core promoter elements based on their defined position immediately upstream of each gene and their nearly universal activation by RNA polymerase II[2,3]. However, the identification of distant acting gene regulatory sequences that direct precise spatial and temporal patterns of expression has been limited despite their established roles in development[4], phenotypic diversity[5] and human disease[6-8]. This has been due to a variety of factors including: the limited number of well-characterized transcription factor binding sites (TFBSs), the degenerate nature of TFBSs and their virtually ubiquitous presence in the genome, as well as difficulties in the development of high throughput functional assays.

Comparative genomic-based approaches have proved to be useful in identifying gene regulatory sequences, primarily on a gene-by-gene basis. These studies involved sequence comparisons of human (or other vertebrate) genomic intervals to orthologous regions from organisms separated by varying evolutionary distances ranging from primates to fish[9-12]. From this work it has been implied that ancient conservation (such as between human and fish) as well as "ultra"-conservation among mammals (sequences at least 200 bp in length that are 100% identical among human/mouse/rat)[1] may be useful indicators of sequences with an increased likelihood of demonstrating gene regulatory activity. These gene-centric investigations, however, have identified only a relatively small number of distant-acting enhancer sequences assayed by a variety of non-uniform methods. In this study, we applied several stringent genomic filters to systematically

identify a large group of putative regulatory sequences and tested their enhancer activity in an *in vivo* transgenic mouse reporter assay[13].

Since one the goals of this work was to assess the validity of a genome-based approach, rather than a gene-centric one, we chose noncoding target sequences based on one of two "extreme" comparative genomic criteria: ancient conservation between human and fugu (separated by ~450 million years of evolution) or ultra-conservation among human-mouse-rat[1]. In total, 167 human DNA fragments were assessed for spatial enhancer activity in a well-established transgenic mouse enhancer assay that links the human conserved fragment to a minimal mouse heat shock promoter fused to a lacZ reporter gene[10,13-16]. We chose to determine tissue-specific reporter gene expression at embryonic day 11.5 (e11.5) since this developmental stage allows for whole-mount staining and whole-embryo visualization. Moreover, at this time-point many of the major tissues and organs have been specified. We also expected this stage to be particularly informative because "extreme" conserved noncoding elements tend to be enriched and clustered near genes expressed during embryonic development[1,12,17,18].

Overall, we found that 29% (24/83) of human-fugu elements alone and 61% (33/54) of human-fugu elements that are also ultraconserved were positive enhancers in this *in vivo* assay (Figure 1; Supplementary Table; the entire dataset including the sequence coordinates, conservation, and whole mount embryo digital imagery can be accessed and queried at the VISTA Enhancer Browser, http://enhancer.lbl.gov). As an example of these data, we present twenty-three elements meeting our selection criteria that were

located in a gene-poor 2.5 Mb stretch bracketing *SALL1*, a transcription factor expressed in early development and mutated in Townes-Brocks syndrome[19] (Figure 2). Seven of the elements flanking *SALL1* directed tissue-specific reporter gene expression in the transgenic *in vivo* assay, recapitulating aspects of *SALL1*'s endogenous expression characteristics at e11.5[20] and further supporting the postulated modular nature of distant acting gene enhancers[21,22]. In addition, we tested 30 ultraconserved noncoding sequences that lacked identifiable conservation with fugu of which 18 (60%) functioned as enhancers, similar to the success rate observed for ultraconserved elements that also have fugu conservation (Figure 1). While the average size of the human fragments tested was 1,270 bp, the positive enhancers overlapped longer human-rodent conserved regions (average length 1,630 bp versus 966 bp; t-test p-value=0.0087; see Methods) and were more conserved among mammals (human-rodent conservation score, t-test p-value=0.0004; see SI Methods) relative to negatives in the assay.

These experimental results reveal the high propensity of extremely conserved human noncoding sequences to behave as gene enhancers *in vivo* and support both ancient human-fish conservation and human-rodent-ultra conservation as highly effective filters to identify such functional elements. The large percentage of elements positive for enhancer activity is particularly surprising considering the single time-point of investigation and the likely possibility that a fraction of the negatives may be enhancers active either earlier or later in development. An important question arising from the significant fraction of ultra and fugu conserved elements functioning as enhancers is whether the tissue-specific enhancer activity that we assess completely explains why

these sequences are so constrained. Overlaying our dataset with results from a recent ChIP-Chip study[23] indicates that at least 7 of the elements reported here (including 4 that are enhancers at e11.5) presumably function as gene silencers in embryonic stem cells. Such data imply that functions in addition to tissue-specific transcriptional activation are embedded in some fraction of extremely conserved noncoding elements, thus potentially contributing to their extreme level of constraint. However, the high efficiency of enhancer identification through this approach nonetheless suggests that tissue-specific transcriptional enhancer activity may be one of the predominant functions of noncoding genomic regions under extreme constraint throughout vertebrate evolution.

We categorized all 75 identified enhancers by their general anatomical patterns of expression using an existing standardized nomenclature[24] (Figure 3). All positive enhancer annotations are based on a minimum of three independent transgenic F0 embryos carrying the same construct and demonstrating the same expression pattern, though the majority (83%) had four or more supporting embryos. We observed reporter gene expression in a variety of anatomical regions, including embryonic structures that are subject to major morphogenetic and remodeling events at e11.5, such as the developing limb, the somites, the heart and the branchial arches (Figure 3). Of the 16 distinct anatomical structures where expression was noted, it was most frequently observed in the central and peripheral nervous system with the most prevalent patterns corresponding to forebrain, midbrain, neural tube, and hindbrain (Figure 3). This bias may be partially explained by the intrinsic complexity of the genetic cascades underlying

vertebrate nervous system development[25] as well as the high percentage of all genes that are expressed in the nervous system.

The majority of the enhancers (50 elements, 66%) directed reproducible expression only to a single anatomical structure at the resolution of whole-mounts. This is consistent with the notion that complex endogenous mRNA expression patterns commonly result from the combined effects of several independent *cis*-regulatory sequences. The remaining one-third (25/75) of the enhancers directed expression to two or more anatomical structures. We speculate that these enhancer elements may be composed of two or more adjacent functional modules that are too tightly linked to each other to be resolved by our comparative approach or that several tissue-specific enhancer activities overlap within a single enhancer element that is utilized in more than one developmental process. Importantly, the enhancer dataset reported here provides a sizeable sequence-based substrate to begin to dissect these possible regulatory mechanisms as well as reagents for further in-depth biological investigation.

To explore if our *in vivo* enhancer dataset could be used to identify sequence features associated with elements driving reporter gene expression in specific anatomical structures, we focused on the forebrain as a test case and selected four of the strongest enhancers identified early on in our survey as a training set. Using a motif finding strategy, we identified six motifs significantly over-represented in these enhancers (see SI Methods). We then scored and ranked all ~3,100 human-fugu conserved noncoding elements in the human genome (Supplemental Table 2) for the statistical over-

representation of these putative forebrain motifs (See SI Methods). The 30 highest-ranking elements included the four known forebrain enhancers that constituted the training set as well as 26 additional elements (Supplemental Table 3). Of these 26 elements, 23 were successfully cloned and tested for *in vivo* enhancer activity in transgenic mice. We observed robust forebrain enhancer activity for 4 of the 23 elements (17%) tested in the transgenic assay system. By comparison, only 4 (5%) of the 77 otherwise uncharacterized human-fugu-conserved elements used to identify the training set were forebrain enhancers (see SI Methods; Figure 4). This preliminary result, while based on a small training set, indicates that a combined comparative and motif-based strategy provided a greater than three-fold enrichment ($p=0.08$, see SI Methods) over comparative-only approaches for the identification of enhancers active in a particular tissue of interest. This initial computational investigation also highlights the need for larger characterized enhancer training sets, the annotation of tissue specificities at high spatial resolution and the development of improved computational methods, which will likely provide a substrate to conclusively establish the predictive power of such approaches.

This study provides quantitative support for previous anecdotal observations that "extreme" evolutionary noncoding conservation is a powerful predictor of mammalian tissue-specific enhancers. Of note, there are at least an additional 5,500 human-fish conserved noncoding sequences in the human genome with similar levels of constraint that are strong candidates for acting as gene enhancers[11]. The efficiency of enhancer identification coupled with the relatively high throughput transgenic assay used here

represents a feasible approach for the generation of a genome-wide experimentally validated enhancer dataset. Such collections are expected to define functional candidate regions as medical sequencing efforts escalate as well as provide a foundation for inferring the network of regulatory interactions among key developmental genes during vertebrate development, analogous to the well-developed efforts in non-vertebrate model systems[21,22]. Regulatory insights derived from these analyses should also enable the creation of modules driving pre-determined expression patterns for various biological applications as well as contribute to an understanding of the vocabulary and grammar of DNA sequences dictating gene expression.

Methods

Identification of conserved noncoding elements and mouse transgenic enhancer assay. Human-fugu conserved noncoding elements with 70% identity, a score of match-mismatch ≥60, and lacking evidence of encoding a protein or being transcribed in mRNA were derived from whole-genome alignments (see SI Methods). The coordinates of ultraconserved elements were retrieved from Bejerano *et al*[1]. Conserved elements were amplified from human genomic DNA by PCR, sequence-validated and transferred into an Hsp68-LacZ reporter vector. Generation of transgenic mice and embryo staining was done as previously described[26] in accordance with protocols approved by the Lawrence Berkeley National Laboratory. For each enhancer fragment, all transgenic embryos exhibiting LacZ-staining were scored and annotated independently by multiple curators.

Motif-finding in a preliminary set of human enhancers and prediction of forebrain enhancers. To find sequence motifs that were associated with forebrain expression, we used a discrete, enumerative motif-finding approach[27]. We identified motifs enriched in the training set of forebrain enhancers relative to three sets of background sequences: 1) random sequences from chromosome 16 (ATTAA and GATTA, which we note are motifs present in previously characterized embryonic forebrain enhancers[28, 29]), 2) a chromosome 16 set of human-fugu fragments (TTNNAAA, CANNGGC and TANNTGA) and 3) a chromosome 16 set of human-fugu sequences that displayed enhancer activity (TTNNTTT) (see SI Methods for details). We then combined information from all the motifs for the prediction of new forebrain enhancers in the

genome by scoring each of 3124 human-mouse-fugu noncoding alignments for the number of conserved (found aligned in human-mouse-fugu) matches to each of the 6 significant 5-mers[18] (see SI Methods for details of scoring procedure). The top 30 fragments are available in Supplementary Table 3.

References

1. Bejerano, G. et al. Ultraconserved elements in the human genome. Science 304, 1321-5 (2004).
2. Roeder, R. G. & Rutter, W. J. Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. Nature 224, 234-7 (1969).
3. Goldberg, M. L. (Stanford University, Stanford, 1979).
4. Stathopoulos, A. & Levine, M. Genomic regulatory networks and animal development. Dev Cell 9, 449-62 (2005).
5. Levine, M. & Tjian, R. Transcription regulation and animal diversity. Nature 424, 147-51 (2003).
6. Emison, E. S. et al. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. Nature 434, 857-63 (2005).
7. Kleinjan, D. A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet 76, 8-32 (2005).
8. Lettice, L. A. et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet 12, 1725-35 (2003).
9. Boffelli, D. et al. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299, 1391-4 (2003).
10. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. Science 302, 413 (2003).
11. Prabhakar, S. et al. Close sequence comparisons are sufficient to identify human cis-regulatory elements. Genome Res (2006).
12. Woolfe, A. et al. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3, e7 (2005).
13. Kothary, R. et al. Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice. Development 105, 707-14 (1989).
14. Rojas, A. et al. Gata4 expression in lateral mesoderm is downstream of BMP4 and is activated directly by Forkhead and GATA transcription factors through a distal enhancer element. Development 132, 3405-17 (2005).
15. Rossant, J., Zirngibl, R., Cado, D., Shago, M. & Giguere, V. Expression of a retinoic acid response element-hsplacZ transgene defines specific domains of transcriptional activity during mouse embryogenesis. Genes Dev 5, 1333-44 (1991).
16. Yamagishi, H. et al. Tbx1 is regulated by tissue-specific forkhead proteins through a common Sonic hedgehog-responsive enhancer. Genes Dev 17, 269-81 (2003).
17. Boffelli, D., Nobrega, M. A. & Rubin, E. M. Comparative genomics at the vertebrate extremes. Nat Rev Genet 5, 456-65 (2004).
18. Ahituv, N., Prabhakar, S., Poulin, F., Rubin, E. M. & Couronne, O. Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny. Hum Mol Genet 14, 3057-63 (2005).

19. Kohlhase, J., Wischermann, A., Reichenbach, H., Froster, U. & Engel, W. Mutations in the SALL1 putative transcription factor gene cause Townes-Brocks syndrome. Nat Genet 18, 81-3 (1998).
20. Buck, A., Kispert, A. & Kohlhase, J. Embryonic expression of the murine homologue of SALL1, the gene mutated in Townes--Brocks syndrome. Mech Dev 104, 143-6 (2001).
21. Carroll, S. B. Evolution at two levels: on genes and form. PLoS Biol 3, e245 (2005).
22. Davidson, E. H. Genomic Regulatory Systems: In Development and Evolution (Academic Press, San Diego, 2001).
23. Lee, T. I. et al. Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 125, 301-13 (2006).
24. Bard, J. L. et al. An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. Mech Dev 74, 111-20 (1998).
25. Gray, P. A. et al. Mouse brain organization revealed through direct genome-scale TF expression analysis. Science 306, 2255-7 (2004).
26. Poulin, F. et al. In vivo characterization of a vertebrate ultraconserved enhancer. Genomics 85, 774-81 (2005).
27. van Helden, J., Andre, B. & Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 281, 827-42 (1998).
28. Kurokawa, D. et al. Regulation of Otx2 expression and its functions in mouse forebrain and midbrain. Development 131, 3319-31 (2004).
29. Zhou, J., Zwicker, J., Szymanski, P., Levine, M. & Tjian, R. TAFII mutations disrupt Dorsal activation in the Drosophila embryo. Proc Natl Acad Sci U S A 95, 13483-8 (1998).

Acknowledgments

Research was conducted at the E.O. Lawrence Berkeley National Laboratory, supported by the Grant # HL066681, Berkeley-PGA, under the Programs for Genomic Application, funded by National Heart, Lung, and Blood Institute, USA, and performed under Department of Energy Contract DE-AC02-05CH11231, University of California.

Correspondence and requests for materials and reprints should be addressed to LAPennacchio@lbl.gov

Figure Legends

Figure 1. A summary of all sequences tested for enhancer activity in transgenic mice. **(A).** A breakdown of the three classes of conserved noncoding sequences assayed: human-fugu alone, human-fugu-ultra, and human-ultra but not fugu. **(B).** The total percentage of positive human enhancers broken down by the same parameters as described in panel A. The total number of elements tested is indicated within panel A while the number of positives is found above the bars of the graph in panel B.

Figure 2. A 3 Mb region of human chromosome 16 enriched for human-fugu noncoding conservation flanking the *SALL1* gene. The coordinates and gene annotations located at the top of the schematic are based on the hg17 assembly at the UCSC Genome Browser (http://genome.ucsc.edu). The middle tracks depict human fragments that were tested in the transgenic mouse enhancer assay and their classification as either "negative" or "positive" refers to their enhancer activity at e11.5. All human elements tested were conserved in the fugu genome and two of these elements were also defined as ultraconserved (denoted by arrowheads). The bottom panel indicates the positive enhancer activities captured through transgenic mouse testing of human-fugu conserved noncoding fragments in this interval.

Figure 3. Grouping of positive expression patterns captured in the transgenic mouse enhancer assay. The total number of elements displaying a given anatomical pattern is depicted by the height of the bars in the chart. A representative transgenic embryo is

provided for each expression pattern.

Figure 4. Application of a forebrain enhancer training set to enrich for forebrain-specific enhancer sequences else where in the human genome. **A.** The four human-fugu chromosome 16 forebrain enhancers used in the training set. **B.** The four positive forebrain enhancers from the 23 human-fugu genome-wide elements predicted to direct forebrain expression based on the training set in A. The UCSC human genome coordinates of the tested fragments(May 2004) and the flanking gene(s) are provided, as well as three representative embryos. Numbers indicate the ratio of forebrain-positive versus the total number of stained embryos.