

## Scanning Human Gene Deserts for Long-Range Enhancers

Marcelo A. Nobrega,<sup>1,2,\*</sup> Ivan Ovcharenko,<sup>1,2,3,\*</sup> Veena Afzal,<sup>1,2</sup> and Edward M. Rubin<sup>1,2,τ</sup>

<sup>1</sup> U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA.

<sup>2</sup> Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

<sup>3</sup> Current Address: Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA

\* These authors contributed equally to this work.

τ To whom correspondence should be addressed: [emrubin@lbl.gov](mailto:emrubin@lbl.gov)

The asymmetric gene distribution in the human genome results in approximately 25% of the genome consisting of gene-poor regions greater than 500 kb, termed gene deserts (1). These large genomic segments have been minimally explored, and their functional significance remains elusive. One category of functional sequences postulated to lie in gene deserts is gene regulatory elements with the ability to modulate gene expression over very long distances (2).

Human *DACHI*, a gene expressed in numerous tissues and involved in the development of brain, limbs and sensory organs (3, 4), spans 430 kb and is bracketed by two gene deserts, 870 kb and 1,330 kb in length. A paucity of regulatory sequences has been identified in the proximity of the *DACHI* promoter (5), suggesting that distal sequences, which could reside anywhere in a sea of sequence greater than 2,630 kb including *DACHI* introns and surrounding gene deserts, are likely responsible for the complex expression characteristics of *DACHI*.

To identify evolutionarily conserved footprints corresponding to putative *DACHI* enhancers, we compared the human *DACHI* sequence and bracketing gene deserts to its orthologous intervals in several vertebrate species (Fig. 1A). Human and mouse sequence comparisons revealed a similar genomic structure within this region and identified 1,098 conserved non-coding sequences (>100bp and >70% identity) in the 2,630 kb targeted interval. To filter these elements into a smaller subset, possibly prioritizing those with a greater likelihood of containing significant biological activity (6), we determined which of the human-mouse conserved sequences were also

present in distant vertebrates, including frog, zebrafish and two pufferfish (*I*). Using both publicly available as well as sequences that we generated, we were able to cover the majority of the orthologous segments in these species (Fig. 1A). Requiring that human-mouse conserved non-coding elements also be present in the majority of distant vertebrates decreased the number of conserved sequences to 32 (Fig. 1B).

To examine the possibility that these sequences, conserved over 1 billion years of parallel evolution time among these vertebrates, might represent enhancers we explored their in vivo ability to drive gene expression utilizing a reporter assay system in transgenic mice. Nine elements were tested, representing a sampling of elements present in the two gene deserts and *DACHI* introns, spread over a 1,530 kb region surrounding the human *DACHI*'s TATA box. Each corresponding human element was individually cloned upstream of a mouse heat shock protein 68 minimal promoter coupled to beta-galactosidase and injected in fertilized mouse oocytes (7). Seven of the nine elements were shown to reproducibly drive beta-galactosidase expression in a distinctive set of tissues in transgenic mice, recapitulating several aspects of *DACHI* endogenous expression (Fig. 1C) (3, 4).

Further support that these elements are functionally linked to *DACHI* expression is provided by the organization of the genes bracketing the orthologous intervals in the vertebrates examined. While the synteny of the orthologous non-coding elements flanking *DACHI* is maintained in mammals and fish, the genes flanking *DACHI* in these vertebrates differ (Fig. 1A). The failure of this chromosomal rearrangement to

disturb the linear relationship between the conserved non-coding elements and *DACHI* further supports a functional relationship between these sequences.

The demonstration that several of the enhancers characterized in this study reside in gene deserts highlight that these regions, occasionally dismissed as but genomic wastelands, can indeed serve as reservoirs for sequence elements containing critically important functions. Moreover, the observation that some of these enhancers can impact human gene expression over extremely large genomic intervals has implications for studies aiming to decipher the regulatory architecture of the human genome, as well as those exploring the functional impact of sequence variation. The size of genomic regions believed to be functionally linked to a particular gene may need to be expanded to take into account the possibility of essential regulatory sequences acting over near megabase distances.

#### **References and Notes:**

1. Materials and Methods are available as supporting material on *Science* Online.
2. L. A. Lettice *et al*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7548 (2002).
3. X. Caubit *et al*, *Dev. Dyn.* **214**, 66 (1999).
4. R. J. Davis *et al*, *Dev. Genes Evol.* **209**, 526 (1999).
5. O. Machon *et al*, *Neurosc.* **112**, 951 (2002).
6. N. Ghanem *et al*, *Genome. Res.* **13**, 533 (2003)
7. R. Kothary *et al*, *Nature* **335**, 435 (1988)

8. We thank I. Plajzer-Frick and J.M. Collier for technical assistance; B. Black for the *hsp68/LacZ* vector. This work was supported by Grant #HL66728, under the NHLBI Programs for Genomic Application, and the U.S. Department of Energy under Contract No. DEAC0376SF00098.

### FIGURE LEGEND:

**Fig. 1. (A)** *DACHI* locus in humans, mice, frog and pufferfish. Lines linking each panel represent positions of orthologous sequences. Genes are represented by their RefSeq name: DAC=DACH1; K1=KLHL1; F=FLJ22624; D=DIS3; P1=PIBF1; G=GPR-18; K=KLF5. **(B)** Sequence conservation plots (alignments were obtained at <http://www-gsd.lbl.gov/vista>). Bars correspond to sequence similarities between human and the species displayed. Blue bars denote exons; red bars denote non-coding sequences. Gradients of red indicate the number of conserved elements within 2,500 bp windows. Asterisks denote elements with no detectable enhancer activity in this developmental stage. **(C)** Transgenic expression results. The distance (in kb) between each element and the human *DACH1* TATA box is given in parenthesis. Expression patterns from representative 12.5 and 13.5 dpc mouse embryos are illustrated. Three or more independent transgenic founders were generated for each element.

FIGURE 1

