Verifying Human Polymorphism and Primate Divergence in HAR1

Sol Katzman, Sofie Salama, Katherine Pollard, Jakob Pedersen, Andrew Kern, Bryan King, David Haussler

Center for Biomolecular Science and Engineering, Howard Hughes Medical Institute, Department of Biomolecular Engineering, University of California Santa Cruz.

INTRODUCTION

We describe Human Accelerated Regions (HARs) as those that are well conserved through recent evolutionary history but which have diverged in the human lineage. To find these, KP first located about 35,000 elements 100bp or longer with 96% sequence identity among chimp, mouse and rat. Assuming for such elements a model like that in Fig.1 (based on a genome wide analysis of conserved regions due to Adam Siepel[1]) Likelihood Ratio Tests were performed to test models in which the human branch has a greater length. KP found about 50 elements having such models with significant p-values. The element with the smallest p-value (< 0.00045) is dubbed HAR1.

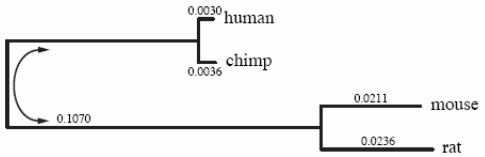

Fig 1: A null model of conserved genomic regions. The HAR1 element has 18 substitutions on the human lineage where only 0.27 would be expected under this null model.

THE HAR1 ELEMENT

The HAR1 element comprises 118bp near the telomere of the q-arm of human chromosome 20. An intriguing feature of this element is that its positive strand is predicted (by the EvoFold program of JP[2]) to form a complex RNA secondary structure. Our lab has sequenced cDNA representing RNA transcripts from both strands, dubbed HAR1F (forward) and HAR1R (reverse, with two splicing isoforms). Analysis shows that none of these are likely to code for proteins. A collaborating lab has found evidence for expression of HAR1F in the developing human brain, leading to the hypothesis that HAR1 plays an important functional role in that process.

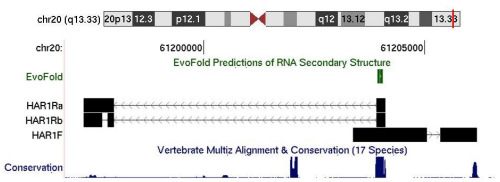

Fig 2: The HAR1 element is found near the telomere of chromosome 20 at a predicted RNA structure. Evidence has been found for RNA transcribed from both strands in this region.

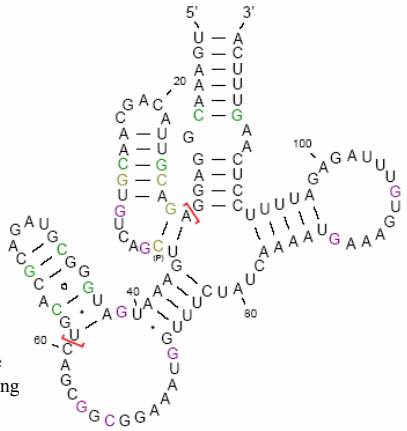

Fig 3: The predicted RNA secondary structure of the HAR1F element in human. The colored bases indicate the human specific changes, with green and yellow highlighting compensatory pairs.

GOALS

We seek to answer some of the following questions: Does the accelerated nature of the HAR1 element in human extend to other parts of the HAR1 gene? Is there evidence from primates besides chimp? If the HAR1 gene in human is special, is it critical to some human function? And is it still evolving?

METHODS

Genomic DNA from 4 primates (chimp, gorilla, orangutan, and crab-eating macaque) was amplified via PCR. The primers for these reactions were designed based on partial genomic sequences available for chimp and rhesus macaque. Sequence reads from clones of several smaller PCR products were concatenated to form sequences covering the human HAR1F exons (Fig.4) For the human study, genomic DNA from 70 individuals was analyzed. For each, a 6500bp region was amplified via PCR and cloned. By sequencing the clones of single PCR products, we acquire a haplotype of the full 6500 bp region for each individual (Fig.5) and can apply statistical tests of evolutionary models for a single locus containing many polymorphic nucleotide sites, as well as attempt to distinguish diseased cases.

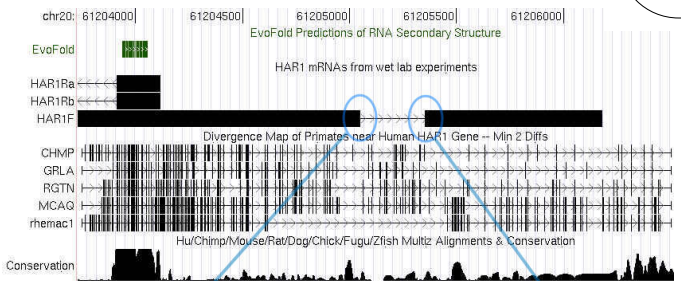

Fig 4: Sequencing results in several primates for the HAR1F transcription region show how human differs strongly (each tic mark is one difference from human reference) from other primates not only at the HAR1 element, but through much of first exon of the HAR1F transcript. The 3' splice site differs only in chimp.

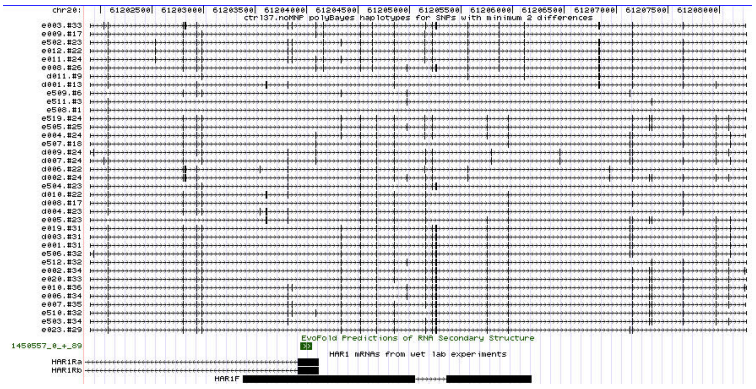

Fig 5: Pattern of SNPs in 37 normal individuals in 6500bp region around HAR1F transcribed element. Additional samples were sequenced of individuals with epilepsy(16) and schizophrenia(17).

RESULTS

The primate sequences confirmed the accelerated nature of the human lineage for the 118bp HAR1 element and most of the first exon of the HAR1F transcript. The 6500bp haplotypes in the 70 human samples contained about 85 polymorphisms. The subset of these consisting of single base substitutions was analyzed with the Tajima D statistic, comparing the frequency spectrum of polymorphisms in the sample to spectra expected under various models of neutral evolution involving population bottlenecks. In no case was the sample value of D statistically significant, so any selective sweep in this region is estimated to have occurred more than one million years ago. Trees built using parsimony of the polymorphisms showed no clustering of the diseased cases.

References

1 Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15, 1034-50 (2005)

2 Pedersen, J. S. et al. Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. PLOS Computational Biology April 2006