Detection of weakly conserved ancestral mammalian regulatory sequences by primate comparisons

Qian-fei Wang[1,2,‡ **], Shyam Prabhakar[1,2, ‡], Sumita Chanan[1], Jan-Fang Cheng[1,2], Edward M. Rubin[1,2], and Dario Boffelli[1,2**]

1 Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720 USA.

2 U.S. Department of Energy Joint Genome Institute, Walnut Creek, California 94598 USA.

‡: These authors contributed equally to this work.

**: Correspondence and requests for materials should be addressed to Q-F.W. (QFWang@lbl.gov) or D.B. (DBoffelli@lbl.gov)

Running Title: Primate-specific Conservation of regulatory Sequences

Abstract

Background

Genomic comparisons between human and distant, non-primate mammals are commonly used to identify *cis*-regulatory elements based on constrained sequence evolution. However, these methods fail to detect functional elements that are too weakly conserved among mammals to distinguish from nonfunctional DNA.

Results

To evaluate a strategy for large scale genome annotation that is complementary to the commonly used distal species comparisons, we explored the potential of deep intra-primate sequence comparisons. We sequenced the orthologs of 558 kb of human genomic sequence, covering multiple loci involved in cholesterol homeostasis, in 6 nonhuman primates. Our analysis identified 6 noncoding DNA elements displaying significant conservation among primates, but undetectable in more distant comparisons. *In vitro* and *in vivo* tests revealed that at least three of these 6 elements have regulatory function. Notably, the mouse orthologs of these three functional human sequences had regulatory activity despite their lack of significant sequence conservation, indicating that they are ancestral mammalian *cis*-regulatory elements. These regulatory elements could be detected even in a smaller set of three primate species including human, rhesus and marmoset.

Conclusions

We have demonstrated that intra-primate sequence comparisons can be used to identify functional modules in large genomic regions, including cis-regulatory elements that are not detectable through comparison with non-mammalian genomes. With the available human and rhesus genomes and marmoset which is being actively sequenced, this strategy can be extended to the whole genome in the near future.

Background

Identifying *cis*-regulatory elements in the human genome, such as promoters and enhancers that regulate gene expression in normal and diseased cells and tissues, is a major challenge of the post-genomic era. Inter-species sequence comparisons have emerged as a major technique for identifying human regulatory elements, particularly comparisons to the sequenced mouse, chicken and fish genomes [1]. However, a significant fraction of empirically defined human regulatory modules are too weakly conserved in other mammalian genomes, such as the mouse, to distinguish from nonfunctional DNA [2], and are completely undetectable in non-mammalian genomes [3, 4]. Identification of such significantly divergent functional sequences will require complementary methods in order to complete the functional annotation of the human genome.

Deep intra-primate sequence comparison, referred to as "phylogenetic shadowing", is a novel alternative to the commonly used distant species comparisons [5]. However, primate shadowing has so far only been applied to the identification of novel *cis*-regulatory elements in short, targeted genomic fragments (≤2.0 kb) [6, 7], due to the lack of sequence data from multiple primates. Thus, it remains to be determined if this approach is useful in identifying otherwise undetectable regulatory regions in an unbiased scans of large genomic loci. Perhaps for this reason, primate shadowing has been almost entirely overlooked as a predictor of regulatory elements.

Here we evaluate the possibility of using deep primate sequence comparisons in large genomic regions (~100 kb) to systematically uncover cis-regulatory elements that are undetectable through mammalian or more distant comparisons. We focused on genes involved in cholesterol metabolism, since this is a physiological process marked by numerous differences between human and distant mammals. In particular, differential regulation of LXRα and its target genes is thought to contribute to inter-species variation in the plasma cholesterol response to dietary cholesterol intake [8]. We evaluated the sensitivity and true positive rate of primate shadowing in identifying known functional sequences in the 8 loci, for which we sequenced a phylogenetically representative panel of primate species. Using a combination of close and distant species comparisons, we then identified 6 human sequences characterized by primate-specific conservation in these 8 gene loci, and tested them for enhancer function *in vitro* and *in vivo*. Finally, we determined if a subset of primate sequences comprising genomes currently available or being sequenced would suffice to identify divergent mammalian regulatory sequences.

Results

Primate comparison identifies known functional sequences in large genomic intervals

To test the power of primate shadowing to identify functional elements in large genomic intervals, we sequenced the primate orthologs of 8 human loci containing *LXRα* and 8 of its target genes: *SREBF1, CYP7A1,* LDL receptor *(LDLR), ABCG5, ABCG8,*

APOE cluster, *APOCIII* cluster, and HMG-CoA reductase *(HMGCR*). The sequenced species comprised 6 anthropoid primates (baboon, colobus, dusky titi, marmoset, owl monkey and squirrel monkey) and one prosimian (lemur). The targeted genomic segments included all exons, introns and flanking intergenic regions of the above mentioned genes, encompassing 558 Kb of human genomic DNA.

We identified sequences evolutionarily conserved among 7 anthropoid primates (the 6 targeted anthropoids plus human) or among all 8 primates (anthropoid plus lemur) using Gumby, an algorithm that detects sequence blocks evolving significantly more slowly than the local neutral rate [3, 9, 10]. Most of the conserved regions overlapped exons (see, for example, Fig. 1). The true-positive rate, defined as the fraction of conserved regions overlapping exons or known regulatory regions, was 80% in the 7-primate comparison (Fig. 2A) and 81% using the 8-primate set (data not shown). The human-dog comparison, which approximately matches the combined branch length of the primate comparison, has a similar true-positive rate of 84% (Fig. 2C). The more distant human-mouse comparison displayed a marginally higher true-positive rate of 90% (Fig. 2B). This is consistent with the theoretical prediction that statistical power increases with the total branch length of the species set [11]. It should be noted, though, that regulatory sequence annotation of the 8 loci we analyzed is probably highly incomplete. Therefore, these true-positive rates are lower bounds; some or all of the "false positives" could eventually be reassigned as true positives upon expansion of the set of sequences annotated as cis-regulatory. Thus, due to incomplete annotation of functional elements, it is not clear if the difference between the primate and human-mouse true positive rates reflects a significant difference in reliability between the two sets of predictions. On average, 64% of the exons in the 8 loci overlapped conserved regions in both the 7-primate and 8-primate comparisons. Similar sensitivity was obtained in human-mouse (65%) and human-dog (71%) comparisons. Thus, primate sequence comparison was approximately equivalent to pairwise human-mouse or human dog analysis in identifying exons.

Phylogenetic shadowing using seven anthropoid primates identifies noncoding sequences with primate-specific conservation

To identify cis-regulatory sequences not detectable in comparisons between human and distant mammals, we searched the 8 gene loci for noncoding sequences highly conserved among primates (p-value ≤ 0.005) but not detectable in human-mouse or human-dog comparisons (p-value > 0.1). Gumby analysis of human and 6 other anthropoid primates identified 6 anthropoid-primate-conserved noncoding regions (Additional data file 4). These sequences were either undetectable (p-value > 0.1) or less significantly conserved (p-value > 0.005) when the prosimian lemur was included in the primate set (data not shown).

To independently confirm the (anthropoid) primate-specific nature of sequence conservation in these 6 regions, we compared their nucleotide substitution rate to that of non-exonic sequences in the same locus. Evolutionarily conserved sequences are defined by a constraint factor (ratio of the substitution rate of test sequences to the non-exonic average) smaller than 1. We found that the 6 primate-conserved sequences had constraint

factors well below 1 among anthropoid primates as expected, and much closer to 1 in the human-mouse and human-dog comparisons (Fig. 3). Finally, none of the 6 sequences overlapped significantly conserved segments identified by the phastCons program [12] in a 17-species alignment of the human genome to (mostly) non-primate mammals and more distant vertebrates [13], which further confirms the primate-specificity of their evolutionary conservation.

Noncoding sequences with primate-specific conservation include three regulatory elements

To explore the potential regulatory function of these primate-conserved elements, we examined their ability to drive reporter gene expression in both a transient transfection assay in human HepG2 cells and an *in vivo* mouse liver gene transfer assay. Since it is possible that the computational prediction only captures part of the entire regulatory module, each human element plus 200-400bp of flanking sequence on either side was cloned upstream of the human promoter of the gene closest to each element and fused to a luciferase reporter gene (See Methods). Therefore, the included flanking sequences may also contribute to the observed regulatory activity. Two elements showed enhancer activity, increasing the expression of a luciferase reporter gene 1.6- to 5-fold in both the human liver cell line HepG2 and *in vivo* in mouse liver, while a third element appeared to be a silencer, suppressing luciferase expression by 50% (Fig. 4, Table 1 and data not shown). While LDLR PS2 showed modest enhancer activity in HepG2 cells (~1.6 fold increase over promoter alone), its activity in 293T cells was much stronger (~5 fold increase over promoter alone), presumably due to availability in these cells of appropriate transcription factors such as SREBFs that are capable to activate LDLR [14]. In an independent assay of transcription potential, both enhancer elements were shown to be DNaseI hypersensitive sites in HepG2 cells (Fig. 4, Table 1 and data not shown), suggesting that the corresponding DNA elements are involved in transcriptional regulation of the endogenous genes. To confirm that the primate orthologs of these identified human regulatory elements are also functional, we cloned the aligned LDLR PS4 sequences from baboon, dusky titi, marmoset and lemur into the luciferase reporter vector and tested their ability to drive reporter gene expression in HepG2 cells. All orthologous non-human primate sequences showed enhancer activity (Additional data file 3).

Regulatory sequences with primate-specific conservation have functional orthologous mammalian counterparts

Since these functional human regulatory elements exhibited primate-specific sequence conservation, we explored whether their functional role is unique to primates. Although human-mouse comparison failed to identify these sequences as constrained (Gumby p-value > 0.1, Fig. 1 and Fig. 3B), we were able to identify the aligned counterparts to the three primate-conserved functional sequences in mouse using the global alignment program MLAGAN [15]. To explore the regulatory function, if any, of these mouse orthologs, we cloned the aligned sequences into the luciferase reporter vector described above and compared their activity to that of the human sequence.

Despite the lack of statistically significant conservation between the rodent and human sequence, all three mouse orthologs exhibited regulatory activity in the same direction to that observed for the human elements (data not shown)**.** Thus, the silencer and the two enhancers identified through primate-specific sequence conservation are ancestral mammalian regulatory elements, rather than newly evolved functional regions specific to primates.

A smaller set of 3 anthropoid primates is sufficient to detect the newly identified regulatory elements

The primate set we used to identify known functional elements and the 3 divergent mammalian regulatory sequences comprises human, baboon, colobus, marmoset, squirrel monkey and owl monkey. However, it is unlikely that all the corresponding genome sequences will be available in the near future. Of the set of most informative primate genomes for comparative analysis [6], the human and rhesus genome sequences are already publicly available, and the marmoset genome is currently being sequenced. We tested whether comparisons among these three species were sufficient to detect functional sequences in the 8 lipid-gene loci. Human-rhesus-marmoset comparison identified 55% of the 160 exons (vs. 7 primates: 64%) with a true-positive rate of 72% (79% for 7 primates) (Fig.2A and D), suggesting that a significant fraction of exons can be detected using a limited number of primates [16]. We subsequently assessed the ability of human-rhesus-marmoset comparison to detect the three newly identified regulatory sequences. As was observed in the comparison of 7 anthropoid primates, both *LDLR* enhancers were highly conserved (p-value <0.005) in the 3-way primate analysis and ranked among the three most conserved noncoding sequences in the 75 kb genomic region (data not shown). The smaller set of 3 primates was also sufficient to detect the silencer in the *SREBF1* locus (noncoding rank: 1), though not as strongly (p-value=0.044, Fig. 1C). The lower statistical significance of the *SREBF1* silencer in the 3-primate analysis relative to the 7-primate analysis is due to the lower combined branch length of the former. These results suggest that availability of the marmoset genome sequence will facilitate genome-wide analysis of primate-specific conservation, and uncover regulatory sequences that are undetectable in distant, non-primate comparisons.

Discussion

Our analysis of over 500Kb of sequence from each of 7 primate species revealed 6 non-coding elements significantly conserved exclusively in primates, of which 3 were found to have gene regulatory activity in a variety of *in vitro* and *in vivo* assays. These 3 regulatory sequences are so weakly conserved in distant, non-primate mammals that none of the three independent methods we tested were able to detect them in mammalian comparisons. However, primate-specific conservation does not imply primate-specific function. Since the mouse orthologs are also functional, the identified sequences appear to be ancestral mammalian regulatory elements, as opposed to newly evolved functional sequences specific to primates. Nonetheless, it is likely that primate sequence comparison could also identify the subset of functional sequences that arose after primates split from distant mammals.

Our results do not of course suggest that primate comparison is optimal for detecting all classes of regulatory sequence. If a human regulatory sequence is constrained in all mammals, for example, then multi-mammal species comparison is clearly preferable to primate comparison, since the mammalian species tree has greater combined branch length, and consequently greater statistical power. However, it is well known that many human regulatory elements show no evidence of constraint in mammalian comparisons [2]. We have demonstrated for the first time that primate comparisons can robustly identify at least some members of this class of "mammal-diverged" human regulatory sequences, even in large (~100 kb) genomic regions.

It is worth noting that the 3 detected regulatory elements displayed only marginal sequence conservation in the prosimian lemur. This result suggests that primate comparisons should be limited to anthropoids (old world monkey and new world monkeys) to sensitively detect divergent mammalian cis-regulatory elements. It is not clear at this point how many additional functional noncoding elements could be detected in the human genome on the basis of primate shadowing, relative to the number of elements already identifiable using the available mammalian genome sequences. However, it is encouraging that, at a conservation p-value threshold of 0.005, primate shadowing expanded the set of predicted noncoding functional elements by 55% (6 elements with primate-specific sequence conservation vs. 11 predicted by human-mouse and human-dog) in the 8 loci examined in this study. Further large-scale studies are required to precisely quantify the value added by multiple-primate analysis.

As a consequence of high sequence identity between humans and great apes, our closest relatives, chimpanzee and gorilla, add very little to the power of primate sequence comparisons. The phylogenetically most informative set of primate species includes Old World monkeys (e.g. rhesus macaque) and New World monkeys (e.g. marmoset), in addition to human [6, 7]. The three functional sequences revealed by 7-primate comparison were also detectable in the three-way human-rhesus-marmoset analysis, albeit less robustly, due to the shorter combined branch length of the 3-way comparisons. Since the human and rhesus genome sequences are already publicly available, and the marmoset genome is currently being sequenced, our results support the feasibility of genome-wide discovery of primate-conserved regulatory elements.

Sequence divergence of the identified regulatory elements between human and distant mammals may reflect functional changes in these sequences. Cis-regulatory elements with primate-specific sequence conservation are therefore potential substrates for determining the molecular basis of primate-specific aspects of gene expression. Previously we described gain of sterol responsiveness in the anthropoid primate LDLR_PS2 enhancer ([17] and Table 1). It is possible that primate-specific sequence conservation of the other two newly identified regulatory elements also reflects qualitative or quantitative expression differences between primates and non-primate mammals, which might be revealed by further in-depth functional characterization and sequence analysis. On the other hand, it is also possible that the lack of significant sequence conservation of some regulatory elements in distant mammals merely reflects the accumulation of compensatory mutations over tens of millions of years, which would retain functional similarity in the absence of significant sequence similarity [18, 19].

Finally, it is possible that short sequence motifs such as transcription factor binding sites within the newly discovered regulatory elements are constrained in all mammals, while the entire elements are significantly conserved only in primates. In one example, we were able to find a conserved functional mammalian AP-4 site in the LDLR-PS2 enhancer ([17] and data not shown).

Conclusion

In summary, our results demonstrate that deep intra-primate sequence comparison can be used to identify functional modules such as exons, enhancers and silencers in large genomic regions. Most importantly, analysis of primate-specific conservation allowed detection of three divergent ancestral cis-regulatory elements, which were not detectable by more distant mammalian comparisons. With the availability of multiple primate genomes, it should be possible to improve the functional annotation of the human genome by uncovering numerous such *cis*-regulatory sequences, some of which potentially contribute to gene expression differences between primates and distant mammals.

Materials and Methods

Sources of sequence and annotation data

Primate BAC clones containing targeted loci were purchased from Children's Hospital Oakland Research Institute in Oakland, California [20]. Draft sequences of baboon, colobus, dusky titi, marmoset, owl monkey, squirrel monkey, and lemur BACs were determined by sequencing ends of 3 Kb subclones to 8-10-fold coverage using BigDye terminators (Applied Biosystems) and assembling reads into contigs with the Phred-Phrap-Consed suite as described previously [21]. All BAC sequences were submitted to GenBank (See Additional data data file 6 for accession numbers). Human, mouse, dog and rhesus sequences were downloaded from the UCSC Genome Bioinformatics website [13]. Based on the human March 2006 assembly hg18, the coordinates for the analyzed human loci are: *SREBF1*: chr17:17653939-17690020; *CYP7A1*: chr8: 59525742-59605413; *LDLR*: chr 19:11054146-11127904; *ABCG5/ABCG8*: chr2:43835998-43966668; *APOE* cluster: chr19:50080832-50149764; *APOCIII* cluster: chr11:116137283-116217351; *HMGCR* chr5:74646522-74714122; *LXRα*: chr11:47231580-47253367. Exon annotations of these regions were obtained from the UCSC Genome Bioinformatics website [13]. The promoter sequence of a gene is defined as the 1 Kb region upstream of the transcription start site. Four enhancers in the *APOE* locus were previously described [22, 23]. These four *APOE* enhancers, together with 15 promoters in the 8 genomic loci, comprise the set of known regulatory regions.

Analysis of sequence conservation

All sequence alignments were carried out using MLAGAN [15]. Aligned sequences were scanned for statistically significant (p-value<=0.1) evolutionarily conserved regions using

Gumby [3, 9, 10]. We defined primate-specific conserved elements as those human sequences that were highly conserved (Gumby p-value <=0.005) among anthropoid primates, but not conserved in more distant mammalian comparisons. Mammalian sequence conservation was defined as: 1) p-value <=0.1 in human-mouse or human-dog comparison, or 2) 70% human-mouse sequence identity over at least 100 bp [24] or 3) significant conservation in an alignment of 17 vertebrate genomes [12].

Evolutionarily conserved regions identified by Gumby were visualized using RankVISTA [25]. Conservation scores in the RankVISTA plots were calculated as the negative logarithm of the Gumby p-value.

The constraint factor of a conserved non-coding sequence (CNS) was defined as the nucleotide substitution rate (summed over all branches of the phylogenetic tree) within the element divided by the local background substitution rate at intronic and intergenic positions in the locus (neutral rate). We estimated substitution rates along each lineage by maximum likelihood using fastDNAml [26].

Plasmid constructs.

The human promoter was cloned in the proper orientation upstream of the luciferase cDNA in the pGL3Basic construct (Promega). The primate-specific elements from human or mouse were PCR cloned into polylinker sites upstream of the promoter of the closest gene (see Additional data file 5 for primer sequences). Each human element includes the primate-conserved sequence plus approximately 200-400bp of flanking sequence. Thus, approximately 1000bp of total sequence is tested in each reporter construct.

Transient-transfection reporter assay.

Cells were grown at 37°C and 5% CO2 in the minimum essential medium (ATCC) (HepG2), or Dulbecco's modified Eagle's medium (ATCC) (293T cells), supplemented with 10% FBS (Hyclone), L-glutamine and penicillin-streptomycin. The cells were grown in 12-well plates ($6x10^4$ cells/well for HepG2, $4x10^4$ cells/well for 293T) and transfected using Fugene (Roche Molecular Biochemicals) following the manufacturer's protocol. Briefly, 500 ng (for HepG2) or 100 ng (for 293T) of each assayed plasmid and 50 ng (for HepG2) or 10 ng (for 293T) pCMVβ (BD Biosciences) were mixed with 1.5 μl Fugene and added to each well. Following 42-48 hours of incubation, cells were harvested and lysed. Activity of luciferase and β-Galactosidase was measured using the Luciferase Assay System (Promega) and the galacto-Light Plus (Applied Biosystems) respectively. Luciferase activity for each sample was normalized to the β-galactosidase assay control. Transfections were carried out in duplicates. All experiments are representative of at least three independent transfections.

Tail vein plasmid DNA transfer assays

Tail vein injection was performed as described by Herweijer and Wolff [27] following the TransIT® *In Vivo* Gene Delivery System Protocol (Mirus Corporation). Six to nine FVB male mice (Charles River Laboratory) at age 7-8 weeks were used for each reporter

gene construct. Ten μg of each reporter construct, along with 2 μg of pCMVβ(BD Biosciences) to correct for delivery efficiency, were injected into each mouse. The entire content of the syringe was delivered in 3-5 seconds. Animals were sacrificed 24 hours later, livers extracted, measured to correct for size, homogenized, and centrifuged for 15 minutes at 4°C, 14,000 rpm. Activity of luciferase and β-Galactosidase was measured as described above. All p-values are from the two-sample Wilcoxon rank-sum (Mann-Whitney) test using STATA (STATA Corporation). All experimental results are representative of two independent plasmid DNA transfer assays.

DNase I-hypersensitive site mapping

DNase I-hypersensitive site mapping was performed as described previously [28]. Briefly, cell pellets were resuspended in DNase I digestion buffer containing 0.5% IGEPAL at 2×10^7 cells/ml buffer. 100-μl aliquots of the resuspended cells were mixed with equal volumes of DNase I buffer containing varying concentrations of DNase I. The DNase I digestion reaction was incubated at 23 °C for 5 min before being stopped with the addition of 8 μl of 0.5 M EDTA and 2 μl of 100 mg/ml RNase A. Following 5 min of RNase A treatment, genomic DNA was isolated using the QiaQuickPCR Kit (Qiagen). 10 μg of each of the DNA samples was digested with appreciate restriction enzyme and resolved in a 1.2% agarose gel by electrophoresis. The DNA from the gel was then transferred onto a nylon membrane. Southern blot was carried out with a radiolabeled DNA probe generated by PCR amplification (see supplementary section for primer sequences), and corresponds to regions ~1000 bp from Gumby predicted elements. Following hybridization and washing, the blot was exposed to Biomax film (Kodak Co.) with intensifying screens at −80 °C for 48 h.

Additional data files:

The following additional data are available with the online version of this paper. Additional data file 1 (Fig.S1_A) and Additional data file 2 (Fig S1_B) show sequence alignments of primate-conserved sequence LDLR_PS2 and SREBF1_PS, respectively. Additional data file 3 (Fig.S2) documents luciferase functional assays, which indicate enhancer activity for orthologous primate *LDLR* PS4 from baboon, dusky titi, marmoset and lemur. Additional data file 4 (Table_S1) documents the coordinates of evolutionarily conserved elements. Additional data file 5 (Table_S2) provides sequences of PCR Primers used for cloning regulatory elements into reporter gene constructs or generating southern blotting probes in detecting DNase I hypersensitive site. Additional data file 6 (Table_S3) is a supplementary table listing GenBank accession numbers for all primate BACs sequenced. Additional data file 7 (Supplemental Figure Legends) contains the figure legends for Fig. S1_A, S1_B and S2.

Additional data files provided with this submission:

Additional data file 1 : FigS1_A.eps
Additional data file 2 : FigS1_B.eps
Additional data file 3 : FigS2.eps
Additional data file 4 : Table_S1
Additional data file 5 : Table_S2.doc
Additional data file 6 : Table_S3.doc

Additional data file 7 : Supplemental Figure Legends.doc

Acknowledgements: We thank A. Moses for discussions and insights on parts of this work, J. Noonan, L. Pennacchio, A. Visel, N. Ahituv, and other Rubin laboratory members for suggestions and criticisms on the manuscript, and B. Kullgren and S. Phouanenavong provided technical assistance for tail vein plasmid DNA transfer assay.

Research was conducted at the E.O. Lawrence Berkeley National Laboratory and at the Joint Genome Institute. This work was supported by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and NIH-NHLBI grant numbers THL007279F and U1HL66681B

References:

1. Miller W, Makova KD, Nekrutenko A, Hardison RC: **Comparative genomics.** *Annu Rev Genomics Hum Genet* 2004, **5:**15-56.
2. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
3. Prabhakar S, Poulin F, Shoukry M, Afzal V, Rubin EM, Couronne O, Pennacchio LA: **Close sequence comparisons are sufficient to identify human cis-regulatory elements.** *Genome Research* 2006, **16:**855-863.
4. Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS: **Conservation of RET regulatory function from human to zebrafish without sequence similarity.** *Science* 2006, **312:**276-279.
5. Boffelli D, Nobrega MA, Rubin EM: **Comparative genomics at the vertebrate extremes.** *Nat Rev Genet* 2004, **5:**456-465.
6. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM: **Phylogenetic shadowing of primate sequences to find functional regions of the human genome.** *Science* 2003, **299:**1391-1394.
7. Soranzo N, Cavalleri GL, Weale ME, Wood NW, Depondt C, Marguerie R, Sisodiya SM, Goldstein DB: **Identifying candidate causal variants responsible for altered activity of the ABCB1 multidrug resistance gene.** *Genome Res* 2004, **14:**1333-1344.
8. Chiang JY: **Bile acid regulation of gene expression: roles of nuclear hormone receptors.** *Endocr Rev* 2002, **23:**443-463.
9. Ahituv N, Prabhakar S, Poulin F, Rubin EM, Couronne O: **Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny.** *Hum Mol Genet* 2005, **14:**3057-3063.
10. Hughes JR, Cheng JF, Ventress N, Prabhakar S, Clark K, Anguita E, De Gobbi M, de Jong P, Rubin E, Higgs DR: **Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences.** *Proc Natl Acad Sci U S A* 2005, **102:**9830-9835.
11. Eddy SR: **A Model of the Statistical Power of Comparative Genome Sequence Analysis.** *PLoS Biology* 2005, **3:**e10.

12. 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.
13. **UCSC web server** [http://genome.ucsc.edu]
14. Hannah VC, Ou J, Luong A, Goldstein JL, Brown MS: **Unsaturated Fatty Acids Down-regulate SREBP Isoforms 1a and 1c by Two Mechanisms in HEK-293 Cells.** *J Biol Chem* 2001, **276:**4365-4372.
15. 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.
16. Ovcharenko I, Boffelli D, Loots GG: **eShadow: a tool for comparing closely related sequences.** *Genome Res* 2004, **14:**1191-1198.
17. Wang QF, Prabhakar S, Wang Q, Moses AM, Chanan S, Brown M, Eisen MB, Cheng JF, Rubin EM, Boffelli D: **Primate-specific evolution of an LDLR enhancer.** *Genome Biol* 2006, **7:**R68.
18. 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.
19. Ludwig MZ, Bergman C, Patel NH, Kreitman M: **Evidence for stabilizing selection in a eukaryotic enhancer element.** *Nature* 2000, **403:**564-567.
20. **BACPAC Resources** [http://bacpac.chori.org]
21. Wang QF, Liu X, O'Connell J, Peng Z, Krauss RM, Rainwater DL, VandeBerg JL, Rubin EM, Cheng JF, Pennacchio LA: **Haplotypes in the APOA1-C3-A4-A5 gene cluster affect plasma lipids in both humans and baboons.** *Hum Mol Genet* 2004, **13:**1049-1056.
22. Allan CM, Taylor S, Taylor JM: **Two hepatic enhancers, HCR.1 and HCR.2, coordinate the liver expression of the entire human apolipoprotein E/C-I/C-IV/C-II gene cluster.** *J Biol Chem* 1997, **272:**29113-29119.
23. Mak PA, Laffitte BA, Desrumaux C, Joseph SB, Curtiss LK, Mangelsdorf DJ, Tontonoz P, Edwards PA: **Regulated expression of the apolipoprotein E/C-I/C-IV/C-II gene cluster in murine and human macrophages. A critical role for nuclear liver X receptors alpha and beta.** *J Biol Chem* 2002, **277:**31900-31908.
24. Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM: **rVista for comparative sequence-based discovery of functional transcription factor binding sites.** *Genome Res* 2002, **12:**832-839.
25. **VISTA web server** [http://pipeline.lbl.gov/]
26. Olsen GJ, Matsuda H, Hagstrom R, Overbeek R: **fastDNAmL: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood.** *Comput Appl Biosci* 1994, **10:**41-48.
27. Herweijer H, Wolff JA: **Progress and prospects: naked DNA gene transfer and therapy.** *Gene Ther* 2003, **10:**453-458.
28. Liang Y, Li XY, Rebar EJ, Li P, Zhou Y, Chen B, Wolffe AP, Case CC: **Activation of vascular endothelial growth factor A transcription in tumorigenic glioblastoma cell lines by an enhancer with cell type-specific DNase I accessibility.** *J Biol Chem* 2002, **277:**20087-20094.

Figure legends

Fig.1 Conservation profiles of a representative region, the *SREBF1* locus, using close (primate) and distant (human-mouse) species comparisons. (A) Seven-primate (human, baboon, colobus, marmoset, dusky titi, owl monkey, and squirrel monkey), (B) human-mouse, and (C) three-primate (human, rhesus, and marmoset) conservation profiles in the SREBF1 locus with flanking genes partially shown. Sequence conservation was calculated using Gumby and visualized using RankVISTA with the human sequence as reference. Vertical bars above the horizontal axis depict evolutionarily conserved sequences, with height indicating the conservation score (-log(conservation p-value), see Methods). Coding exons (dark blue) and UTRs (magenta) are marked below the horizontal axis. Vertical bars that overlap coding exons or UTRs are colored light blue, while nonoverlapping bars are colored red. The arrow denotes SREBF1_PS, a noncoding element conserved in primates (p-value ≤ 0.005) but not in the mouse (p-value > 0.1).

Fig. 2 Primate comparisons identify known functional elements and conserved noncoding sequences in genomic intervals encompassing *LXRα, SREBF1*, *CYP7A1, LDLR, ABCG5, ABCG8, APOE* cluster, *APOCIII* cluster, and *HMGCR*. The number of evolutionarily conserved sequences (p-value$\leq$0.1) overlapping exons, previously known regulatory elements and unannotated regions (new predictions) in the 8 loci are shown for the following species sets: A) 7 anthropoid primates, B) Human-mouse, C) Human-dog and D) Human-rhesus-marmoset. Percentages were calculated by dividing the number of conserved sequences of each type by the total number of conserved elements (x100).

Fig.3: Evolutionary conservation of 6 primate-conserved sequence in anthropoid primates, but not between human and mouse or dog. (A) Sequence alignment of a representative primate-conserved sequence LDLR_PS4. Similar alignments for LDLR_PS2 and SREBF1_PS are provided as Additional data files 1 and 2. (B) The constraint factor of a sequence element is defined as the nucleotide substitution rate (total branch length of the phylogenetic tree) within the element relative to the background noncoding rate in the aligned sequences. Constraint factors in the anthropoid primate comparisons (black bars) are consistently well below one (dashed line). Human-mouse (dotted) and human-dog (white) constraint factor ranges of the 6 sequences are broader, mostly exceeding one at the upper limit. Error bars indicate 95% confidence intervals.

Fig. 4 Functional assays indicate enhancer activity for a representative primate-conserved element, human *LDLR* PS4. Luciferase assay analysis of (A) transient transfections into human HepG2 cells and (B) plasmid DNA transfer into mouse liver. The luciferase reporter constructs tested are either the LDLR promoter alone (promoter), or the promoter in combination with the human LDLR PS4 (+ PS4). Fold increase over the empty vector is shown. Error bars indicate standard deviation (A). Each triangle in (B) represents luciferase activity in an individual mouse. Red bars denote the median activity of each construct. Luciferase activity is reported in arbitrary units. (C) DNaseI hypersensitive site mapping around LDLR PS4 region in human liver cell line HepG2. Vertical arrow indicates the lane with internal size marker that was generated by enzyme digestion of the LDLR PS4 sequence. The hypersensitive site (HS) is indicated by a

horizontal arrow. Co-migration of the internal size marker with the HS localizes the HS to LDLR PS4 sequence.

Table 1: Functional characterization of noncoding elements significantly conserved only in primates. Human elements with primate-specific conservation were tested for their ability to drive reporter gene expression *in vitro* in HepG2 cells and *in vivo* in mouse liver. The genomic regions containing primate-conserved elements were also examined for the presence of DNaseI hypersensitive sites (DNaseI HS) in HepG2 cells. Enhancer or silencer strength is shown as fold increase or decrease relative to the promoter alone in luciferase assays.

Primate Specific Element	***In vitro* (HepG2)**		***In vivo* in Mice**
	DNaseI HS	Reporter Transfection	Gene Transfer
LDLR_PS1	No	No activity	No activity
LDLR_PS2	Yes	Enhancer* (~5.1 Fold)	Enhancer (~5.5 Fold)
LDLR_PS3	No	No activity	No activity
LDLR_PS4	Yes	Enhancer (~3.7 Fold)	Enhancer (~4.2 Fold)
SREBF1_PS	No	Silencer (~2.4 Fold)	Silencer (~1.8 Fold)
CYP7A1_PS	No	No activity	No activity

*: in 293T cells.

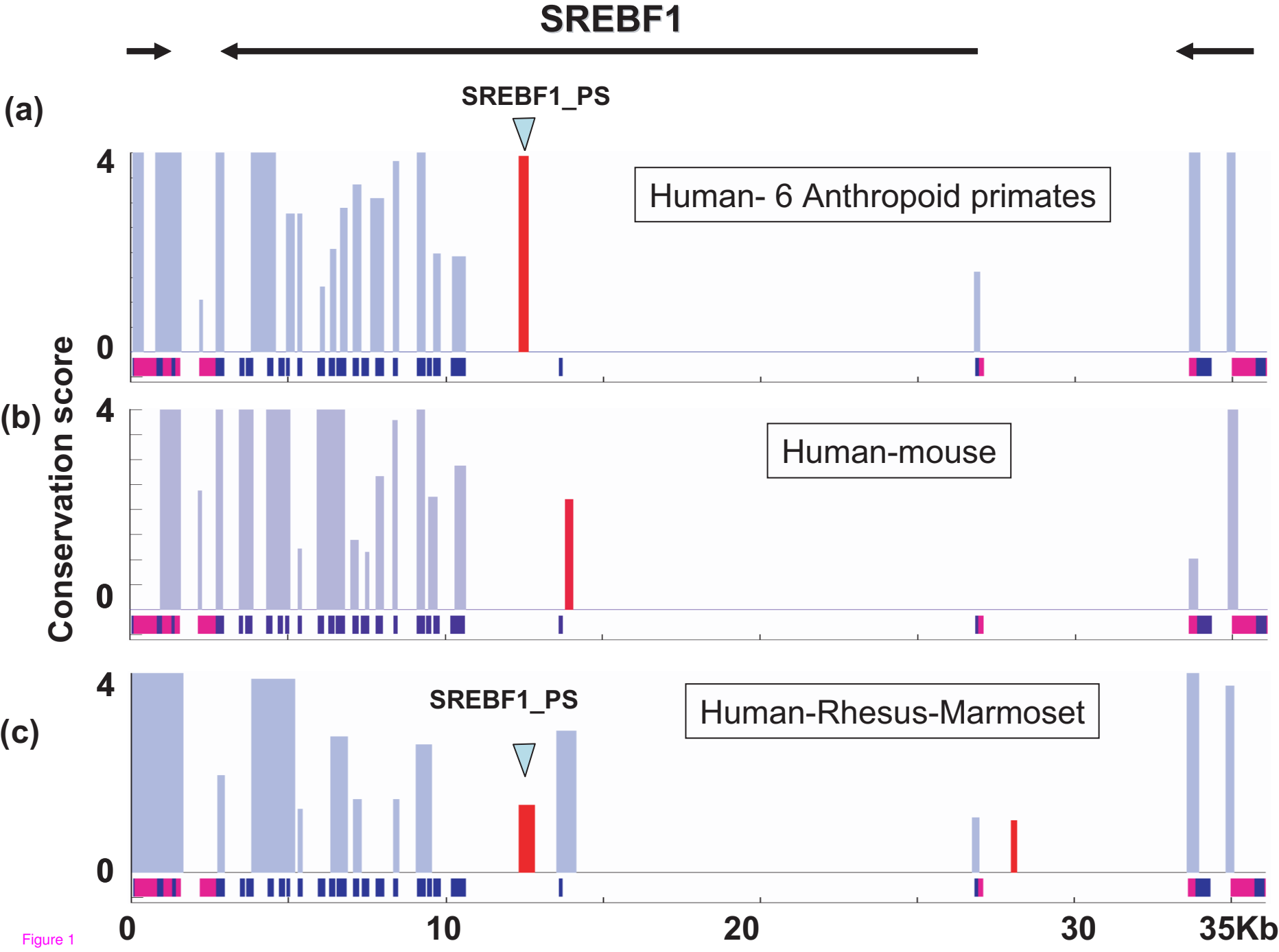
SREBF1
(a)
SREBF1_PS
4
0
Human- 6 Anthropoid primates
Conservation score
(b)
4
0
Human-mouse
(c)
4
SREBF1_PS
Human-Rhesus-Marmoset
0
0
10
20
30
35Kb

Figure 1

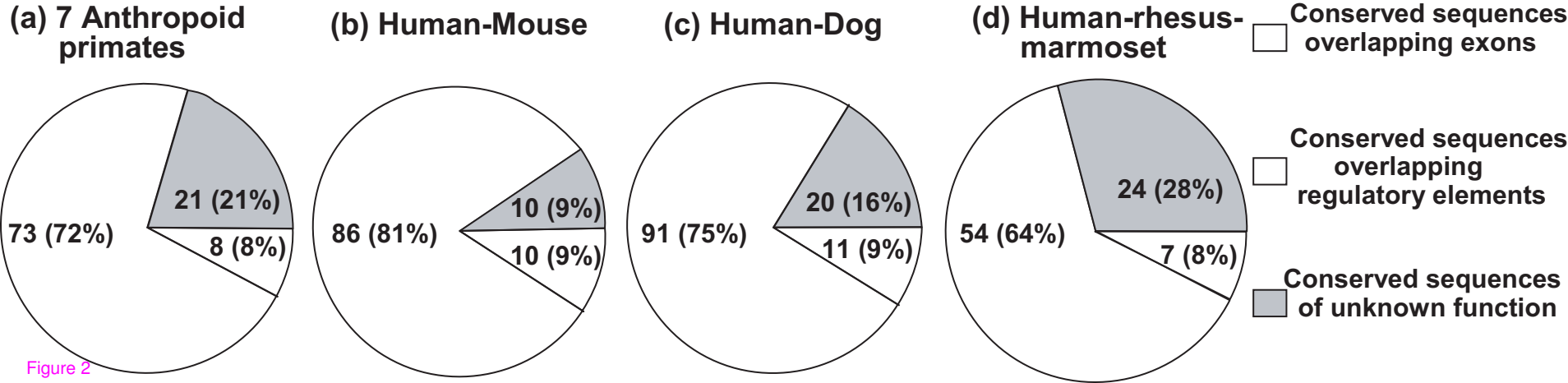
(a) 7 Anthropoid primates
21 (21%)
73 (72%)
8 (8%)
(b) Human-Mouse
10 (9%)
86 (81%)
10 (9%)
(c) Human-Dog
20 (16%)
91 (75%)
11 (9%)
(d) Human-rhesus-marmoset
24 (28%)
54 (64%)
7 (8%)
Conserved sequences overlapping exons
Conserved sequences overlapping regulatory elements
Conserved sequences of unknown function

Figure 2

(a)

```
human          1 CTGGGTTCATAACCTGGCTCTGCCACTGACTCGCTGGGTGACAGGCACCTCAGAGCCTCAGTTTCCCCACCTGACAATGG
baboon         1 CTGGGTTCATATCCTGGCTCTGCCACTGACTCGCTGGGTGACGGGCACCTCAGAGCCTCAGTTTCTCCACCCGACAATGG
duskytiti      1 CTGGGTTCATATCCCGGTTCTGCCACTGACTCGCTGGGTGACGGGCACCTCAGAGCCTCAGTTTCCCCACCCGACAATGG
marmoset       1 CTGGGTTCATATCCCGGTTCTGCCACTGACTCGCTGGGTGACAGGCACCTCAGAGCCTCAGTTTCTCCACCCGACAATGG
squirrlmonk    1 CTGGGTTCAT-TCCCGGGTCTGCCACTGACTCGC-GTGTGACAGGCACCTCAGAGCCTCAGTTTCCCC-CCCCACAATGG

lemur          1 CTGGGTTCATATACTGGCTGTGCCGCTGATTCGCTCAGTGACGGGCATCTTAGAGCCTCAGTTTCCCCATCT--------
mouse          1 CCTGGCTCCGACCCTGGCTCTGCTG-TAGTTTTCTGTGTGACGTGCACTTCAGAGCCTC-GTTAGCCCAG----------
dog            1 ------------------------------------------------CCTTGGAGCCTCAGTTTCCCCAGCTGAAGCTGG

human         81 AGACAAAGCTAATCTCCCCCTCCCCAGGGGCTCTGGAAGTGGGGCAGGATGGGGCTGCGCAGGCGCTCGGAGACAAAGGC
baboon        81 AGACAAAGCTAATCTCCCCCTCCCCAGGGGCTCTGGAAGTGGGATGGGATGGGGCTGCGCAGGCGCCCGGAGACAAAGGC
duskytiti     81 AGACAGAGCTGATC-CCCCCTCCCCAGGGGCTCTGGAAGTGGGGCTGGATGGGGCTGCACAGGCGCCCGGAGACAAAGGC
marmoset      81 AGACAAAGCTGATCTCCCCCTCCCCAGGGGCTCTGGAAGTGGGGCAGGATGGGGCTGGCAGGCACCCGGAGACAAAGGC
squirrlmonk   78 AGACAAAGCTGATCTCCCCCTCCTCAGGGCCTCTGGAAGTGGGGCGGGATGGGGCTGCGCAGGCGCCCGGAGACAAAGGC

lemur         73 -GACAATGCTAATCTCCG---CCTCCTGGGCTGTGTGAGAGGGGTGGGCTGGGGCTCTACACAGGCCTGGAGACAAAGGC
mouse         69 ---------TAATT-CCACCTTCT-----GCCA---------GACTGGTGGGACCTCTACCC--------------AGGC
dog           34 GGCCAGGGCTAATCCCCAGCTCCC---AGGCTC------AGAGCCCGTGCGGGGCCCTGCACAGGCCTGGAGAAGAAGGC

human        161 AGGGCCTGTCATCTTTCCTGCGTCCACGGGGTGGCACTTCCTTCCTGCCTTCCGCCTTTTGTTTTGGGACGCTTCCAAAA
baboon       161 AGGGCCTGTCATCTTTCCCGCGTCCACAGGGTGGCACTTCCTTCCTGCCTTCCGCCTTTTGTTTTGGAAAGCTTCCAAAA
duskytiti    160 AGGGCTTGTCATCTTTCCCCTGTCCACAGGGTGGCGCTTCCTTCCTGCCTTCCGCCTTTTGTTTTGGGAAGCTTCCAAAA
marmoset     161 AGGGCCTGTCATCTTTCCCCTGTCCACAGGGTGGCACTTCCTTCCTGCCTTCCGCCTTTTGTTTTGGGAAGCTTCCAAAA
squirrlmonk  158 AGGGCCTGTCATCTTTCCCCTGTCCACAGGGTGGCACTTCCTTCCTGCCTTCCGCCTTTTGTTTTGGGAAGCTTTCAAAA

lemur        149 AGGGCCTGTCATCTTTCCTGCGTCCACAGGGTGGCACTTCCTTCCTGCCTTCCGCCTTTTGTGTTGGGAAGCTTCCAAAA
mouse        111 -----CTGCCATCTCCTCTGAATCCACGGGGTGGCACATCCGTCCTTCCTTCTGCCCTTTGTTTGGGGCAAGAGCCAAAC
dog          105 GGGGACTGTCATGTTCGCTGTGTCCACAGGGTGGCACTTCCTTCCTTCCTTCCGCCCTTTGTCTTGGGAAGGTTCCAAAA

human        241 CGCCCCTGGGAGGGAAAACTGAGCAGCCCACACAGGAAGCGTCCTGGAGCCTGCACACAGGCGCTCGATAATTGCTCGAT
baboon       241 CGCCCCTGGGAGGGAAAATGGAGCAGCCCACACAGGAAGCGTCCCGGAGCCTGCACACATGCGCTCGATAATTGCTTGAT
duskytiti    240 CGCCCCTGGGAGGGAAAACCGAGCCGCCCACACAGGAAGCGGCCCCCGGCCTGCACACAGGCGCTCAATAATTGCTTGAT
marmoset     241 CGCCCCTGGGAGGGAAAACCGAGCAGCCCACACAGGAAGCGGCCCCCGGCCGGCACACAGGCGCTCAATAATTGCTTGAT
squirrlmonk  238 CGCCCCTGGGAGGGAAAACCGAGCAGCCCACACAGGAAGTGGCCCCCGGCCGGCACACAGGCGCTCAATAATTGCTTGAT

lemur        229 CGCCTCTGGGGA-GTGAAACTGAAGAGCCCACACAGGAAGCG--CTGGGGCCTGCACACG-GCGCTCAATAATCGCTTGCT
mouse        186 TGCCTCAGAGAGGGGAGAAGCAAGTGG-----CCTGGAAAGCCC-----GTGGCTTCAC--AGGCCTTCAAGGACTGCCTCCC
dog          185 TGCCTCTGGGGGGTGAAGGCTGAACGG-GCGCCCGGGAAGCGCGCCGGAG----------GCCGCTCGATAATTGGTTGAT

human        321 TGACGAAATTGGTGCTCAACCAAGTGGCAAACAGGATAAGCGGGCTCAGATGGCCAGGAAAACGGGA
baboon       321 TGACGAAATTGGTGCTCAACCAAGTGGCAAACAGGATAAGTGGGCTCAGATGGCCAGGAAAACGGGA
duskytiti    320 TGACGAGATCGGTGCTCAACCGAGTGGCAAACAGGATAAGCAGGCTCGGATGGCCAGGAAAACGGGA
marmoset     321 TGACGAAATCTGTGCTCAACCGAGTGGCAAACAGGATAAGCAGGCTCGGGTGGCCGGGAAAACGGGA
squirrlmonk  318 TGACGAAATCGGTGCTCAACCGAGTGGCAAACAGGATAAGCAGGCTCGGGTGGCCAGGAAAACGGGA

lemur        305 GGAATAAATCAGCGCTCAACCGAGTGGCAAATGGGATAAATGATCTCGGGTGGCCTGGAACACGGCA
mouse        256 TGAGTCA--TGGTGCTGGCATGTGTGGCTAACAGCATGGTGGAGCTTCCACGTCCTGGCAGACTGGA
dog          254 TGACTGATTCCAAGCTCAGCCGCGCGGCAAACGGGCCAAATGAGCGTAAATGGCCCCGGCAAACCGAA
```

(b)

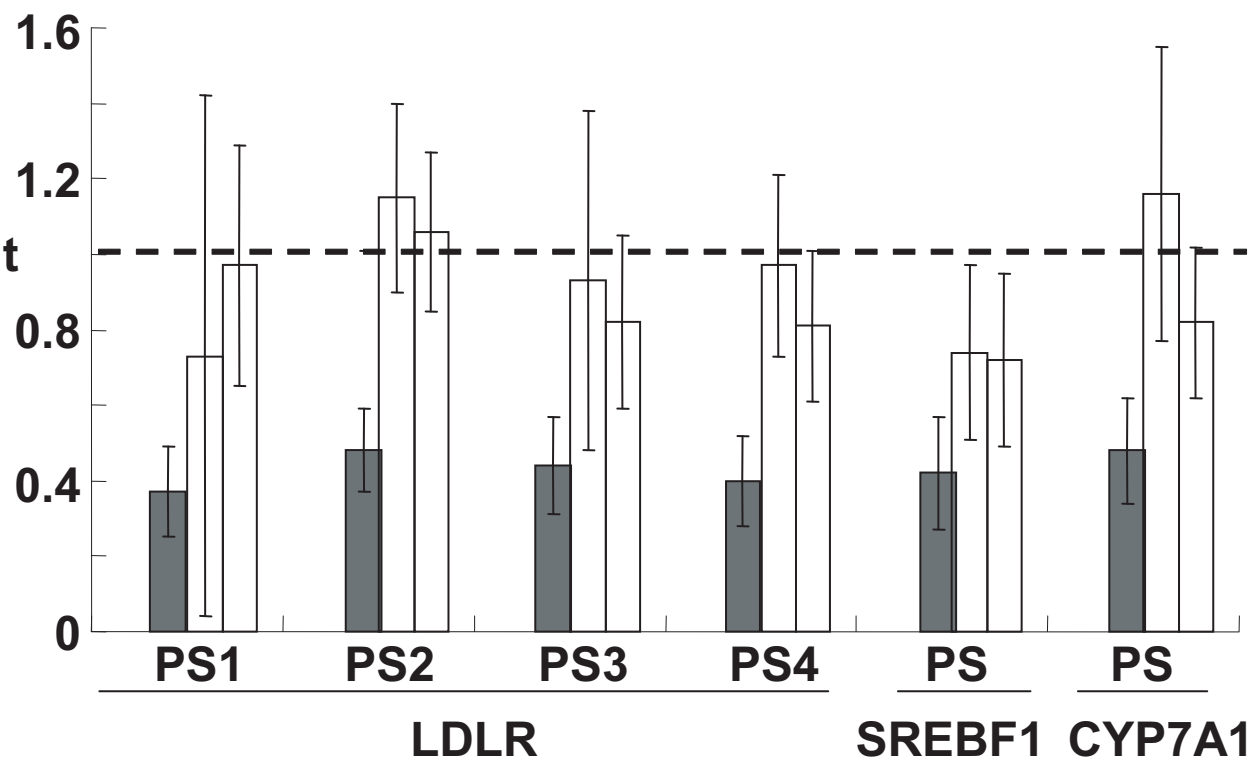

Figure 3

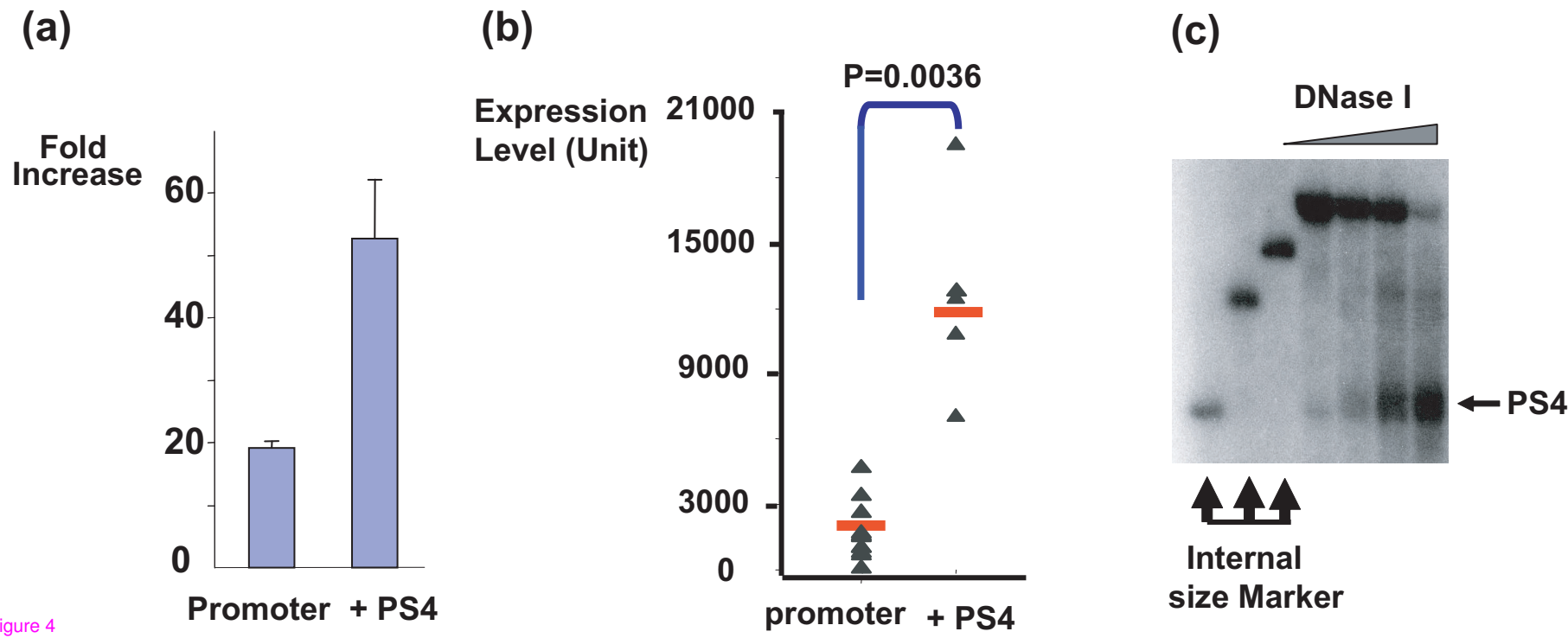
(a)
Fold Increase
60
40
20
0
Promoter
+ PS4
(b)
P=0.0036
Expression Level (Unit)
21000
15000
9000
3000
0
promoter
+ PS4
(c)
DNase I
PS4
Internal size Marker

Figure 4

Additional files provided with this submission:

Additional file 7 : supplemental figure legends.doc : 24Kb
http://genomebiology.com/imedia/1015104136126515/sup7.DOC
Additional file 6 : table_s3.doc : 26Kb
http://genomebiology.com/imedia/5759296101265158/sup6.DOC
Additional file 5 : table_s2.doc : 30Kb
http://genomebiology.com/imedia/7031697612651583/sup5.DOC
Additional file 4 : table_s1.doc : 25Kb
http://genomebiology.com/imedia/1288214415126515/sup4.DOC
Additional file 3 : figs2.eps : 1031Kb
http://genomebiology.com/imedia/1568111686126515/sup3.EPS
Additional file 2 : figs1_b.eps : 2379Kb
http://genomebiology.com/imedia/8857247891265158/sup2.EPS
Additional file 1 : figs1_a.eps : 4138Kb
http://genomebiology.com/imedia/9839284181265158/sup1.EPS