Recent natural selection in human noncoding sequences

Heather A. Lawson, Joel Martin, David C. King, Belinda Giardine, Webb Miller, Ross C. Hardison

Center for Comparative Genomics and Bioinformatics, Huck Institutes of Life Sciences, and Departments of Anthropology, Biology, Computer Science and Engineering, Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802

Corresponding author:
Ross Hardison
Email: rch8@psu.edu
Phone: 814-863-0113
FAX: 814-863-7024

Abstract

Genomic DNA sequences associated with human-specific functions should show a signature of positive selection, whereas those needed in all primates should show a signature of negative selection. One test for natural selection (either positive or negative) that has been applied to protein-coding regions requires a ratio of intraspecies polymorphism to interspecies divergence that deviates significantly from the neutral expectation. By applying this test to noncoding regions of human DNA, using ancestral repeats in tiled windows as a model for neutral DNA, we found that about 1.0 to 1.6% of the human genome deviates significantly from neutrality based on comparisons to rhesus (diverging about 23 million years ago) or chimpanzee (diverging about 6 million years ago), respectively. Most of these non-neutral windows show a signal for positive selection. Many of them overlap with genes such as *FOXP2* whose coding regions were previously implicated as being targets of positive selection in humans, and others overlap with novel candidates for selection, such as the nerve growth regulator *NEGR1*. These noncoding regions that are candidates for targets of selection could harbor regulatory regions, such as a predicted *cis*-regulatory region in the tumor suppressor *PDSS2*. Non-neutral intronic regions in the *FOXP2* gene have protein-binding and modification properties consistent with such a role. Thus novel candidates for selection can be identified in potential *cis*-regulatory modules.

Only a small fraction of mammalian genomes, roughly 5% (*1-4*), is thought to carry out functions conserved in all mammals. This fraction should be subject to negative (purifying) selection and thus can be detected as sequences showing significantly less change than neutral DNA over the span of mammalian evolution. Other functional sequences have undergone relatively recent and rapid change, and hence appear to be under positive selection. Recent papers (*5, 6*) have identified protein-coding genes showing evidence of such selection on the human lineage. Most current tests for recent selection have been applied only to protein-coding regions, e.g. McDonald-Kreitman (*7*), HKA (*8*), and a variety of tests based on rates of substitution in synonymous and nonsynonymous sites (*9*). However, functional sequences under selection include both protein-coding genes and noncoding sequences, such as *cis*-regulatory modules (CRMs). A dramatic example of presumptive positive selection are the highly accelerated regions of the human genome (*10, 11*). Also, a set of conserved noncoding sequences was tested for accelerated evolution by examining human-specific substitutions in interspecies comparisons (*12*). Here we apply a modification of the McDonald-Kreitman test to identify longer genomic intervals showing evidence of recent positive and negative selection in human noncoding regions. In addition to divergence data between human and either chimpanzee or rhesus, this test uses human polymorphism data and is applied to the entire genome.

The McDonald-Kreitman test (*7*) compares the counts of polymorphisms (within species) and fixed differences (between species) at nonsynonymous sites within coding regions with the polymorphism and divergence counts at synonymous sites, which are assumed to be neutral. In general, nonfunctional sites should show a ratio of polymorphism to divergence counts (r_{pd}) indistinguishable from the r_{pd} at neutral sites (*13*). Sites that are under positive selection will diverge more between species than the neutral sites, and thus they will have an r_{pd} smaller than expected from neutrality. Conversely, sites that are under negative selection will diverge less between species, and they will show an r_{pd} greater than expected from neutrality. Other factors, such as population expansion, can also affect the r_{pd} (*14*). The ratio r_{pd} in a feature class, e.g. genes or windows of DNA, can be compared with the r_{pd} for neutral sites in the vicinity to find regions that deviate significantly from neutrality by this measure (*7, 13, 15*).

We implemented a variation of the McDonald-Kreitman test (MKAR test) on tiled 10kb windows of genomic human DNA using ancestral repeats (ARs) in the windows as a neutral model (*1, 2, 20-22*). The human polymorphism data was from dbSNP version 126 (*16*), filtered to retain validated single nucleotide polymorphsims (SNPs) in nuclear DNA and to remove SNPs that map to multiple locations. Divergence was computed over both the human-chimpanzee and the human-rhesus time scales, approximately 6 million years ago (Mya) and 23Mya, respectively (*17*). The protein-coding portions of exons were masked, so divergence and polymorphism counts are only made in noncoding regions. A 2x2 contingency table contained polymorphism and divergence counts in non-AR and AR sites for each window; deviation from neutrality was evaluated for significance by a chi-square test. By comparing counts in the nonAR sites to those in AR sites in each window, we capture the local variation in neutral evolutionary rates (*1, 18*). To correct for multiple testing, a q-value was computed from the empirical distribution of chi-square p-values and used to estimate the false discovery rate (FDR) (*19*) (Fig. 1A). Results can be viewed and analyzed as a custom track on the UCSC Genome Browser (*20*) at http://hgwdev-giardine.cse.ucsc.edu/cgi-bin/hgTracks (track name="McDonald-Kreitman AR"). Similar results are obtained with windows of larger size, but with smaller sizes, most windows do not have sufficient counts for the test.

Using a p-value of 0.01 from the chi-square test as a threshold for significance, approximately 1.6% of the windows in the human genome deviate from neutrality when divergence is computed for human versus chimpanzee (Fig. 1A). A similar analysis using human divergence from rhesus showed that approximately 1.0% of the windows are significantly non-neutral. Many of these non-neutral windows (26%) are the same in the two tests, but a larger fraction is found only in the comparisons with chimpanzee (50%) or rhesus (24%) (Fig. 2). Thus different regions can show this signature of selection over distinct evolutionary times, approximately 6 million (chimp) and 23 million (rhesus) years ago.

The direction of apparent selection is indicated by the neutrality index (NI), which is the ratio of r_{pd} for a feature compared to r_{pd} for the local neutral DNA. Values greater than 1 are associated with negative selection and values less than 1 are associated with positive selection (*21*). Most windows have an NI close to 1 (Fig. 1B), but for those that significantly deviate from neutrality, more appear to be under positive selection than negative (Fig. 1B, Fig. 2A and B). In contrast, most protein-coding genes surmised to be under selection are subject to negative (purifying) selection (*5*). This suggests that noncoding sequences may be under adaptive selection more frequently than protein-coding sequences.

When interpreting any test for selection using polymorphism and divergence data, one must be aware of the limitations in the input data. Some of the data in dbSNP126 come from datasets with known biases toward more frequent alleles (*22, 23*). This bias will apply equally to both AR and nonAR sites, so the possible effect on the MKAR results may not be large. One approach to evaluating the magnitude of the effect of this ascertainment bias is to compare MKAR results obtained using dbSNP 126 data with results obtained using a polymorphism dataset that should be less biased toward frequent alleles. This dataset combined resequenced HapMap data (*22*) with HapMap Phase II data in the 10 resequenced ENCODE (*24*) regions. For the 417 10kb windows that could be tested in these resequenced ENCODE regions, the *p*-value for deviation of windows from neutrality correlated between the MKAR tests with an *r* of 0.535 (using divergence from chimp). A similar comparison using divergence from rhesus gave an *r* of 0.538. The NI values correlated as well, with *r* = 0.531 and 0.539 for divergence from chimpanzee and rhesus, respectively (p-value < 2.2×10^{-16} in both cases). Thus even with known biases in the current polymorphism data, the test is fairly robust. However, it will be important to continue to evaluate signals for recent selection as more polymorphism data become available, especially for less biased data. The divergence data also have limitations, largely in the quality of the assemblies of the comparison species, but also in the contribution of non-orthologous regions to the alignments (for example, see Fig. 3). As genome assemblies and alignments improve, it will be important to re-run these tests.

The MKAR test detects signals suggestive of positive selection in regions containing genes and members of gene families previously demonstrated to be under selection (Table 1 and Supplementary Tables A and B). With divergence computed against either chimp or rhesus, these include *CRB1*, *CYB5R4*, *DMD*, *FOXP2*, *NEURL*, TRPV6, *ZNF493*, and *ZP3* (*5, 25, 26*), as well as members of tumor suppressor families, the olfactory receptor family, and the cadherins. This family of morphoregulatory genes are differentially expressed in both the developing and the mature brain (*27*), and they have been implicated in other recent studies as the target of positive selection in noncoding regions in human (*12*). Examination of Gene Ontology (*28*) classification using the program GOStat (*29*) on the genes overlapping significantly non-neutral windows showed an overrepresentation of genes associated with nine terms. Three of the more specific terms were “nervous system development,” “visual perception” and “cell communication” (Supplementary Tables C,D,E and F).

The genome-wide MKAR test also indicates that positive selection is occurring in several genes not previously implicated in adaptive evolution, and this selection appears to be in *cis*-regulatory modules (CRMs). Based on human-chimpanzee divergence, 896 genes overlapped 2867 10kb windows significantly suggestive of positive selection by the MKAR test and NI<1 (Supplementary Table A). Using human-rhesus divergence, we find 654 genes overlapping 1772 windows with a signature of positive selection (Supplementary Table B). After applying a multiple test correction, 75 and 34 genes (divergence from chimpanzee and rhesus, respectively) were found overlapping windows whose MKAR result has a false discovery rate (FDR) of 5% or lower (Table 1 and Supplementary Tables G and H). One example of a gene that includes a region significantly deviant from neutrality based on human polymorphism and divergence from rhesus is *NEGR1*, which encodes a regulator of nerve growth (Fig. 3). A region close to the first exon is significantly non-neutral in tests using comparisons with either chimpanzee or rhesus, and with estimated FDRs of 16% and 2%, respectively (Fig. 3A and 3B). The polymorphism counts are similar for both nonAR and AR sites, but the divergence from rhesus is much greater in the nonAR sites (Fig. 3C), which is consistent with positive selection (NI of 0.21). The excess of divergence in the nonAR sites leads to a pronounced difference in the r_{pd} values (Fig. 3C) and thus a highly significant result in the MKAR test.

Furthermore, the non-neutral window in *NEGR1* has properties associated with a CRM (Fig. 3B). It is in an intron, and some segments of the window have a high value for regulatory potential, which is a score based on machine-learning of strong and weak signals in alignments of known CRMs (*30*). These data suggest that a regulatory region within this segment has undergone substantial nucleotide changes since the divergence of human from rhesus, perhaps leading to adaptive changes in the regulation of expression of *NEGR1*. The MKAR test on windows does not resolve the actual sequence that is under selection but rather it shows that the 10kb window has a signature associated with selection. The target of selection could be the putative CRM or it could be the region surrounding it.

This example also illustrates the value of being able to view the MKAR results in the context of other information on the UCSC Genome Browser (*20*). Whole genome alignments include matches between both orthologous (inferred to be derived from the same ancestral sequence) and some paralogous (produced by duplications) sequences in the compared species. Tests for selection are designed for comparisons of orthologous sequences, and comparisons of paralogous sequences can lead to artifactually high divergence counts. By viewing the alignment net tracks (*31*) along with the MKAR test results, we see that part of the non-neutral window is aligned with paralogous sequence in the rhesus (Fig. 3B). When the test is repeated after masking the duplicated region, the results are still significant for this window (Fig. 3C). However, it is prudent to examine any region of interest for possible artifacts using information in genome browsers.

FOXP2 also shows evidence of selection in a potential CRM (Fig. 4). Two windows are significantly deviant from neutrality by the MKAR test (chi-square p-values< 0.01). The one in intron 2 is significant in comparisons with both species and it has an NI associated with positive selection. Furthermore, data from the ENCODE consortium (*32*) shows evidence of hyperacetylation of histone H4, occupancy by transcription-related proteins (RARA, Myc, E2F4, and BAF170), and depletion of nucleosomes, as indicated on the FAIRE (formaldehyde assisted isolation of regulatory elements, (*33*)) and DNase hypersensitivity (DHS) tracks. The other non-neutral segment, located in intron 5, shows a strong signal for negative selection (NI of 11.7 for comparison with chimpanzee). This segment shows evidence of DNA methylation in HepG2 cells and methylation of histone H3 on lysine 27 (Fig. 4). These are modifications often associated with gene repression, and it is possible that this non-neutral segment is a target for down regulation of the *FOXP2* gene. The transcription factor FOXP2 is the target of positive selection in humans and has been implicated in language development (*34*). These new analyses suggest the general locations (to within 10kb) of CRMs that are also under selection, which could lead to lineage-specific adjustments in the level of the FOXP2 protein in certain tissues.

In order to find particular functional regions, it is desirable to test for selection on individual DNA segments. This has been done in two recent studies utilizing whole genome interspecies comparisons, both focused on DNA segments that show evidence of negative selection in mammals. Human accelerated regions, or HARs, are the most rapidly changing DNA segments in humans (*10, 11*). Also, a subset of conserved noncoding sequences (CNSs) was found to undergo accelerated evolution in humans (*12*). The MKAR test looks at all human sequences that align with other primates, and thus it is more inclusive. Also, it explicitly takes into account local neutral evolutionary rate variations captured by the AR polymorphism and divergence data, which the other tests do not. However, it is limited when examining individual regions, because the polymorphism counts in isolated members of a feature set (such as HARs or CNSs) are too small to provide statistical power. Thus MKAR tests on individual short regions are not meaningful, but it is instructive to examine overlaps with these other tests. Of the 202 HARs (average size 173bp), 171 overlap 10kb windows with polymorphism and divergence counts for the MKAR test, but none of the windows deviate significantly from neutrality. Of the 992 CNSs (average size 268bp) showing acceleration in the human lineage, 841 overlap with 10kb windows with counts for the MKAR test, and 13 (1.5%, divergence reckoned with chimp) and 5 (0.6%, divergence reckoned with rhesus) of these windows deviate significantly from neutrality (listed in Supplementary Table I). These percentages are similar to those seen for all windows genome-wide, and thus it appears that our non-neutral windows are not enriched in the features that other studies have found to be under accelerated evolution in humans. We expect that any individual feature must contribute a strong signal for non-neutrality to drive a 10kb window to significance. Of the 699 genes in the vicinity of accelerated CNSs (*12*), 157 also overlap with a window significantly deviating from neutrality by the MKAR test (listed in Supplementary Tables J and K). Interestingly, in only two of these cases does the gene's associated CNS also overlap an MKAR non-neutral window. This suggests that in the vast majority of cases, the MKAR test is finding a signal in a region different from the accelerated CNS, and illustrates the novelty of the MKAR approach.

We also searched for overlap between a genome-wide collection of stringently defined, candidate CRMs and significantly non-neutral windows by the MKAR test. The candidate CRMs are DNA segments predicted to be CRMs by two independent methods. The first method found 282,639 DNA segments of at least 200bp with consistently high regulatory potential scores (*30, 35*). The second method finds 118,402 conserved clusters of matches to transcription factor binding sites called PReMods (*36*). We investigated the set of 92,269 predicted CRMs in both sets, called PRPs. The PRP intervals overlapping windows with an FDR of 5% or less are shown in Table 2 and Supplementary Tables L and M). One of these is in an intronic segment of the tumor suppressor gene *PDSS2* (Fig. 5). Both the *p*-value and the FDR for this segment are below 10^{-15} in comparisons with both chimpanzee and rhesus. The NI is also very low (0.02 to 0.03), associated with positive selection. Thus this intronic DNA segment is a strong candidate for containing a CRM that has been under selection for a human-specific change in pattern of expression. Another striking example is chr13:99,889,901-99,890,101, found in the intronic region of *PCCA*, variations in which are associated with the enzyme deficiency propionic acidemia.

The results of this MKAR test can be used to find noncoding regions throughout the human genome that are candidate targets of natural selection over the past 23Myr (comparisons with rhesus) or 6Myr (comparisons with chimpanzee). These are potentially functional noncoding sequences, and some may be CRMs. The examples presented here show evidence consistent with roles in gene regulation, including stringent prediction through largely independent methods (PRPs) and biochemical evidence of protein binding and chromatin modification from the ENCODE data. By leveraging polymorphism counts and recent divergence, the MKAR test explores a phylogenetic span, and likely some functional regions, different from those investigated by studies of conservation among different mammalian orders or conservation from mammals to other vertebrates, such as birds and fish. Thus the results of this test should be a useful addition to resources of comparative genomics aimed at finding functional DNA sequences. These results also point to candidates for noncoding regions that play a role in human-specific properties. Experimental tests of these candidates should be a fruitful area for further investigation.

Methods

Datasets examined

Divergence data for the MKAR test was generated from three-way multiple alignments of human-chimp-macaque computed on the March 2006 human assembly (hg18), the March 2006 chimp assembly (panTro2), and the January 2006 macaque assembly (rheMac2) using MULTIZ (*37*). Polymorphism data for the MKAR test was obtained from dbSNP build 126 (*16*) downloaded from the UCSC Genome Table Browser (*20*) and filtered for genomic SNPs with a known validation status. Ancestral repeats were obtained from RepeatMasker (*38*) output filtered for repeats having 90% or greater alignment among human, chimp and macaque, with primate-specific repeats (those with a milliDiv count greater than 180, e.g. 18% divergence from consensus) filtered out. We also removed MER121 repeat family members, demonstrated to violate neutral expectations (*39*). All coding exon start and stop positions were downloaded from the UCSC Genome Table Browser RefSeq Genes (*40*) genes and gene predictions track for human genome release hg18. ENCODE resequenced regions and HapMap resequenced data combined with HapMap Phase II derived allele frequency data for the Yoruban population were obtained from the UCSC Genome Table Browser. Coordinates for the HapMap resequenced SNPS were converted from human genome release 17 to human genome release 18 using the Batch Coordinate Conversion (liftOver) utility on the UCSC Genome Browser.

The set of predicted *cis*-regulatory modules, PRPs, used in this study was derived by overlapping predicted regulatory module (PReMods, *n* = 118,402) (*36*) start and stop positions with intervals having high Regulatory Potential (RP, *n* = 282,639) scores as determined by ESPERR (*30, 35*). PReMods are computationally determined using information from Transfac (*41*) position weight matrices for vertebrate transcription factors and clustering of putative binding sites. PReMod’s are available for download at (http://genomequebec.mcgill.ca/PReMod). RP sites, deliberately ignorant of Transfac elements, are computationally determined by applying learning algorithms to 7-species multiple alignments. RP sites are available for download at (http://www.bx.psu.edu/projects/esperr). All coordinates were converted from human genome release 17 to human genome release 18 using the Batch Coordinate Conversion (liftOver) utility on the UCSC Genome Browser.

MKAR Test

Windows of size 10kb were tiled across the genome. A 2X2 contingency table was generated for each window by apportioning SNP counts and divergence counts within a window into AR and non-AR categories (Supplementary Table N). A SNP count is registered when a SNP is present in an aligned position. A divergence count is registered when aligned bases differ and no SNP is present. An AR count is registered when either a SNP or diverged base fall in an ancestral repeat site within a window. A non-AR site is registered if either a SNP or a diverged base fall outside an AR site within a window. Gaps or regions that do not align in either species pair (human-chimp or human-macaque), or windows with a zero count in any member of the 2X2 contingency table were not counted. The significance of the table's deviation from neutrality was determined by the standard chi-square test (Fisher's exact test if any member of the table was less than 5) as implemented in R (*42*). The FDR multiple tests correction procedure was applied to the empirical distribution of chi-square p-values as implemented in R, using the qvalue library and the bootstrap method of estimating the proportion of true null hypotheses (*19*).

Genes, predicted cis-regulatory modules, HARs, and accelerated conserved noncoding sequences overlapping windows

10kb windows were joined with a list of RefSeq gene names and accession numbers using the Galaxy inner-join function (*43*). 10kb windows were joined with a list of overlapping PRPs (described in main text) in the same manner. The set of HARs was obtained from Pollard et al., 2006 (*5, 6*) supplementary materials. The set of accelerated conserved noncoding sequences was obtained from Prabhakar et al. 2006 (*12*) supplementary materials. Intervals were lifted from hg17 to hg18 and joined with 10kb windows.

Gene Ontology analysis

The probability that the Gene Ontology categories overlapping statistically significant 10kb windows were from a random sampling of all Gene Ontology categories was determined by GOStat (*29*) using a Benjamini correction, which controls the false discovery rate (*44*).

Evaluation of ascertainment bias

The mMK test was run using the HapMap resequenced data in the ENCODE resequenced regions. Windows corresponding to those generated using dbSNP126 were paired and the Pearson's correlation, as implemented in R, was compared between chi-square p-values and NI values.

Acknowledgements

This work was supported by NIH grants from NHGRI (HG002238, WM), NIDDK (DK65806, RH), by Tobacco Settlement Funds from the Pennsylvania Department of Health, and by the Huck Institutes of Life Sciences, Pennsylvania State University.

References

1. R. H. Waterston *et al.*, *Nature* **420**, 520 (2002).
2. F. Chiaromonte *et al.*, *Cold Spring Harb Symp Quant Biol* **68**, 245 (2003).
3. K. Lindblad-Toh *et al.*, *Nature* **438**, 803 (2005).
4. A. Siepel *et al.*, *Genome Res* **15**, 1034 (2005).
5. C. D. Bustamante *et al.*, *Nature* **437**, 1153 (2005).
6. R. Nielsen *et al.*, *PLoS Biol* **3**, e170 (2005).
7. J. H. McDonald, M. Kreitman, *Nature* **351**, 652 (1991).
8. R. R. Hudson, M. Kreitman, M. Aguade, *Genetics* **116**, 153 (1987).

9. Z. Yang, J. P. Bielawski, *Trends in Ecology and Evolution* **15**, 496 (2000).

10. K. S. Pollard *et al.*, *PLoS Genet* **2** (2006).

11. K. S. Pollard *et al.*, *Nature* **443**, 167 (2006).

12. S. Prabhakar *et al.*, *Science* **314**, 786 (2006).

13. H. Akashi, *Genetics* **139**, 1067 (1995).

14. Y. X. Fu, *Genetics* **147**, 915 (1997).

15. S. A. Sawyer, D. L. Hartl, *Genetics* **132**, 1161 (1992).

16. S. T. Sherry *et al.*, *Nucleic Acids Res* **29**, 308 (2001).

17. C. B. Stewart, T. R. Disotell, *Curr Biol* **8**, R582 (1998).

18. R. C. Hardison *et al.*, *Genome Res* **13**, 13 (2003).

19. J. D. Storey, *Journal of the Royal Statistical Society Series B-Statistical Methodology* **64**, 479 (2002).

20. W. J. Kent *et al.*, *Genome Res* **12**, 996 (2002).

21. D. M. Rand, L. M. Kann, *Mol Biol Evol* **13**, 735 (1996).

22. The International HapMap Consortium, *Nature* **437**, 1299 (2005).

23. A. G. Clark *et al.*, *Genome Res* **15**, 1496 (2005).

24. The ENCODE Project Consortium, *Science* **306**, 636 (2004).

25. S. Biswas, J. M. Akey, *Trends Genet* **22**, 437 (2006).

26. P. C. Sabeti *et al.*, *Science* **312**, 1614 (2006).

27. C. Redies, *Progress in Neurobiology* **61**, 611 (2000).

28. M. Ashburner *et al.*, *Nat Genet* **25**, 25 (2000).

29. T. Beissbarth, T. P. Speed, *Bioinformatics* **20**, 1464 (2004).

30. J. Taylor *et al.*, *Genome Res* **16**, 1596 (2006).

31. W. J. Kent *et al.*, *Proc Natl Acad Sci U S A* **100**, 11484 (2003).

32. The ENCODE Project Consortium, *Nature* ((submitted) 2007).

33. P. L. Nagy *et al.*, *Proc Natl Acad Sci U S A* **100**, 6364 (2003).

34. W. Enard *et al.*, *Nature* **418**, 869 (2002).

35. D. Kolbe *et al.*, *Genome Res* **14**, 700 (2004).

36. M. Blanchette *et al.*, *Genome Res* **16**, 656 (2006).

37. M. Blanchette *et al.*, *Genome Res* **14**, 708 (2004).

38. A. F. A. Smit, R. Hubley, P. Green, *RepeatMasker Open-3.0*, <http://repeatmasker.org >, (1996-2004).

39. M. Kamal, X. Xie, E. S. Lander, *Proc Natl Acad Sci U S A* **103**, 2740 (2006).

40. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res* **33**, D501 (2005).

41. V. Matys *et al.*, *Nucleic Acids Res* **31**, 374 (2003).

42. R Development Core Team, *R: A language and environment for statistical computing*, <http://www.r-project.org/ >, (2005).

43. B. Giardine *et al.*, *Genome Res* **15**, 1451 (2005).

44. B. Efron, *Journal of the American Statistical Association* **99**, 96 (2004).

Figure Legends

Figure 1. Distributions of p-values for deviation from neutrality by the MKAR test (A) and neutrality index (B). In (A), the distribution of negative logs of the p-value determined by a chi-square test for all windows is plotted in the larger graph, and the distribution for the windows passing a p-value threshold of 0.01 is plotted in the inset. The thresholds for 5% and 1% FDR are also shown in the inset. In (B), the distribution of neutrality index (NI) values for all windows is shown in the upper graph, and the distribution for the windows that pass a p-value threshold of 0.01 is shown in the lower graph.

Figure 2. Significantly non-neutral windows found over different evolutionary timeframes. The Venn diagram shows the numbers of windows significantly deviating from neutrality that are found using divergence from chimpanzee (blue circle) or from rhesus (red circle). The yellow disks in each part of the diagram show the number of windows inferred to be under negative selection, and the others are inferred to be under positive selection.

Figure 3. Windows deviating from neutrality in the *NEGR1* gene. Panel (A) shows the data for 10kb windows throughout the gene, which is shown as a series of boxes (exons) connected by a line (introns) with arrows showing the direction of transcription.

The next three lines plot the negative logarithm of the p-value determined in an MKAR chi-square test, the neutrality index (NI), and the negative logarithm of the false discovery rate (FDR), all determined using divergence from chimpanzee. The next three lines plot data for the same functions, using divergence from rhesus. Panel (B) focuses on the highly significant window in intron 1 close to the start of transcription, showing only the MKAR chi-square p-value tracks. The regulatory potential based on alignments of seven species (7X Reg Potential), alignment nets of human with rhesus (Rhesus Net) and chimpanzee (Chimp Net), and human self alignments are also shown. The color of each block on the alignment nets indicates the chromosome with the aligning sequence in the second species. Brown indicates chromosome 1, whereas the yellow and green blocks in the rhesus net mean that the aligning segments are from chromosome 9 and chromosome 3, respectively, and thus are not orthologous to the human sequence. This portion of human is similar to other human DNA, as shown by the gray rectangles on the Human Self Alignments track. Panel (C) gives the polymorphism and divergence counts in the nonAR and AR sites in the window, the ratio of polymorphism to divergence (r_{pd}), and statistics on signficance, all determined using divergence from rhesus. Panels A and B were made from displays generated on the UCSC Genome Browser (*20*).

Figure 4. Windows deviating from neutrality in the *FOXP2* gene, and ENCODE annotations. The conventions for the display are the same as in Figure 3, except that the MKAR test results are determined using divergence from chimpanzee. In addition, selected tracks of ENCODE data are presented. These are trimethylation of lysine 27 of histone H3 (H3K27me3), binding of the retinoic acid receptor alpha (RARA), hyperacetylation of histone H4 in K562 cells (H4ac K562), binding by Myc in fibroblasts (Myc Fb), binding by E2F4 in fibroblasts (E2F4 Fb), binding by BAF170, DNA methylation in HepG2 cells (Meth HepG2), formaldehyde assisted isolation of regulatory regions in HeLa cells (FAIRE) and DNase sensitivity in SKnSH cells (DHS SKnSH). Counts and results of significance tests for the two boxed windows are given in the lower portion of the figure.

Figure 5. Windows deviating from neutrality in the *PDSS2* tumor suppressor gene, with PRPs. The conventions for the display are the same as in Figure 3; the MKAR test results shown were determined using divergence from rhesus. PRPs are DNA segments predicted to be *cis*-regulatory modules by two different approaches: clusters of conserved matches to transcription factor binding motifs (*36*) and high regulatory potential (*30, 35*).

Table 1. Selected genes overlapping 10kb windows significantly suggestive of positive selection

				Divergence from					
				Chimpanzee			Rhesus		.
Gene	Chrom	Window Start	Window Stop	NI	chi-square p-value	FDR q-value	NI	chi-square p-value	FDR q-value
CRB1	chr1	195530000	195540000	0.16	9.0E-03	5.9E-01	0.67	7.1E-01	1
CYB5R4	chr6	84630000	84640000	0.10	2.7E-03	3.8E-01	0.084	8.5E-04	4.2E-01
DMD	chrX	31560000	31570000	0.023	8.3E-03	5.8E-01	0.12	6.7E-02	1
FOXP2	chr7	113890000	113900000	0.15	2.2E-04	1.0E-01	0.24	3.9E-03	7.6E-01
NEGR1	chr1	72510000	72520000	0.15	4.6E-04	1.6E-01	0.15	3.5E-06	9.4E-03
NEURL	chr10	105280000	105290000	0.15	7.0E-03	5.5E-01	0.13	5.3E-03	8.3E-01
TRPV6	chr7	142290000	142300000	0.28	7.0E-02	9.7E-01	0.12	5.5E-03	8.4E-01
ZP3	chr7	75880000	75890000	0.021	3.1E-04	1.3E-01	0.31	7.3E-02	1
*OR5P3**	chr11	7800000	7810000	0.32	1.2E-02	6.5E-01	0.28	4.6E-04	3.0E-01
*CDH4***	chr20	59390000	59400000	0.31	4.3E-03	4.6E-01	0.25	5.5E-03	8.4E-01
*PDSS2****	chr6	107710000	107720000	0.15	4.0E-03	4.4E-01	0.033	0	0
*PLAGL1****	chr6	144340000	144350000	0.14	2.8E-06	4.4E-03	0.16	8.0E-07	2.7E-03

* indicates a member of the olfactory receptor family

** indicates a member of the cadherin family

*** indicates a tumor suppressor

Table 2. Selected PRP intervals overlapping windows significantly suggestive of positive selection

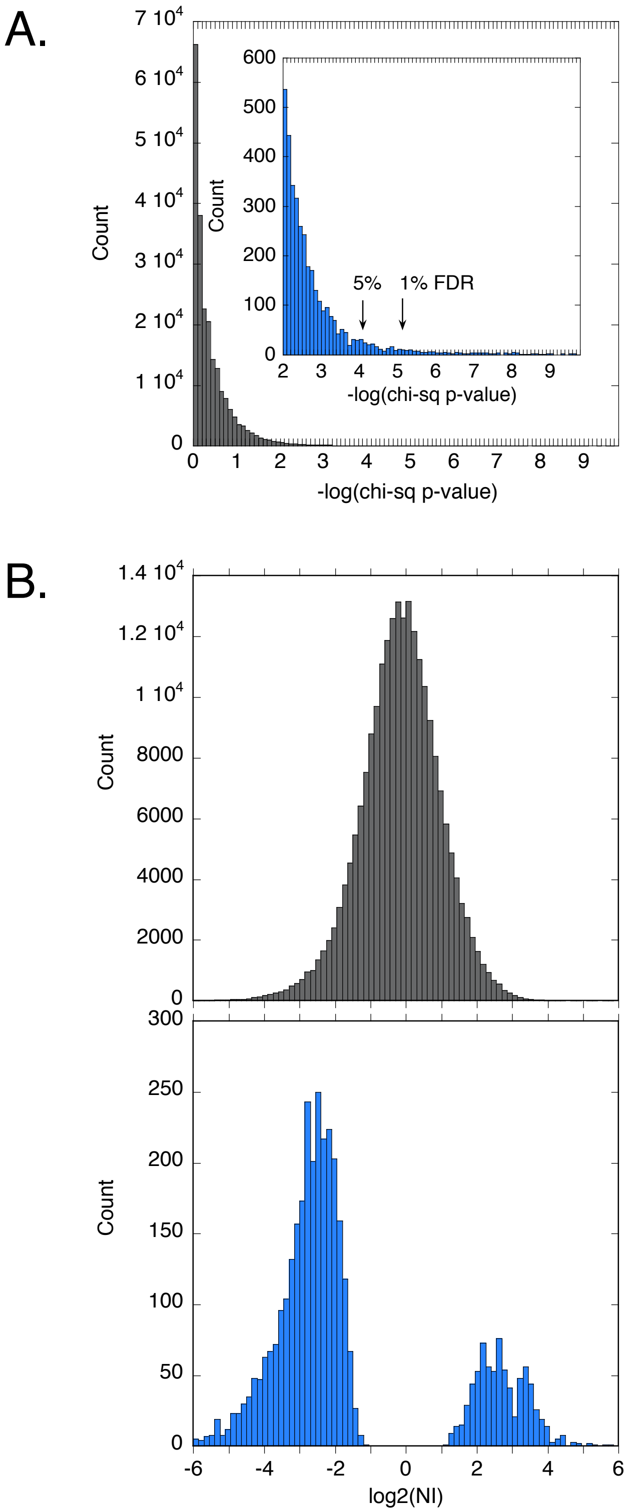
Chromosome	Interval Start	Interval Stop	Nearest Gene
chr1	27924716[C]	27925061	*FAM76A*
chr3	160621015[C]	160621273	*SCHIP1*
chr6	107720808*	107721505	*PDSS2*
chr6	14249988[R]	14250061	*CD83*
chr6	147870401[R]	147871443	*LOC389432*
chr8	143628479*	143628742	*BAI1*
chr9	18614889[C]	18614941	*ADAMTSL1*
chr9	23694695[R]	23694766	*ELAVL2*
chr10	127894951[C]	127895222	*ADAM12*
chr10	31161100*	31161155	*ZNF438*
chr11	106833541[C]	106833874	*CWF19L2*
chr13	99889901*	99890101	*PCCA*
chr16	7406121[R]	7406273	*A2BP1*
chr17	33153549[C]	33154254	*TCF2*
chr17	60020606[C]	60020707	*SMURF2*
chr18	51342916[R]	51343094	*TCF4*
chr19	44081672[C]	44081752	*NFKBIB*
chr19	44082268[C]	44082422	*SIRT2*
chr20	42504721[C]	42505004	*HNF4A*

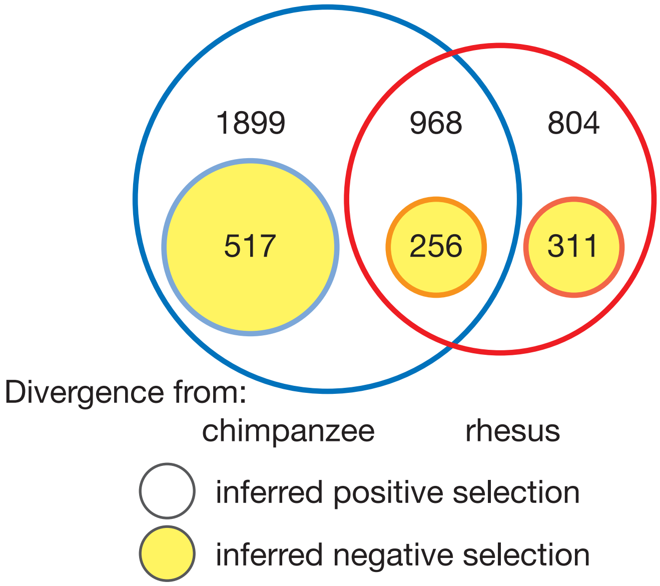
Windows with an FDR *q*-value ≤ 0.05 and NI<1 that overlap PRPs are listed.

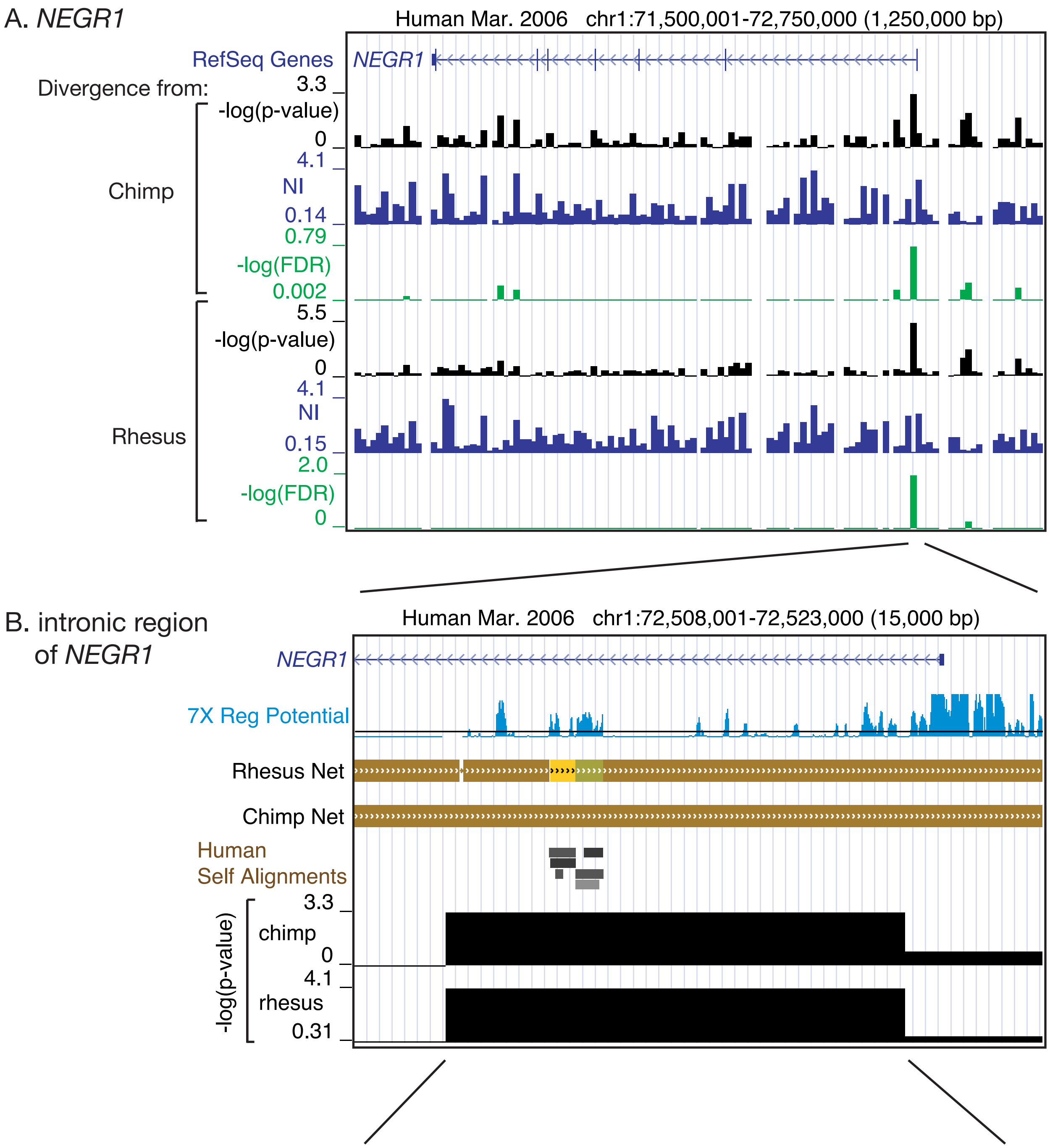
[C] indicates an interval found only with divergence reckoned with chimpanzee

[R] indicates an interval found only with divergence reckoned with rhesus

* indicates an interval found with divergence reckoned with both chimpanzee and rhesus


A.
7 10^4
6 10^4
5 10^4
4 10^4
3 10^4
2 10^4
1 10^4
0
Count
0 1 2 3 4 5 6 7 8 9
-log(chi-sq p-value)
600
500
400
300
200
100
0
Count
5%
1% FDR
2 3 4 5 6 7 8 9
-log(chi-sq p-value)
B.
1.4 10^4
1.2 10^4
1 10^4
8000
6000
4000
2000
0
Count
300
250
200
150
100
50
0
Count
-6 -4 -2 0 2 4 6
log2(NI)


1899
968
804
517
256
311
Divergence from:
chimpanzee
rhesus
inferred positive selection
inferred negative selection



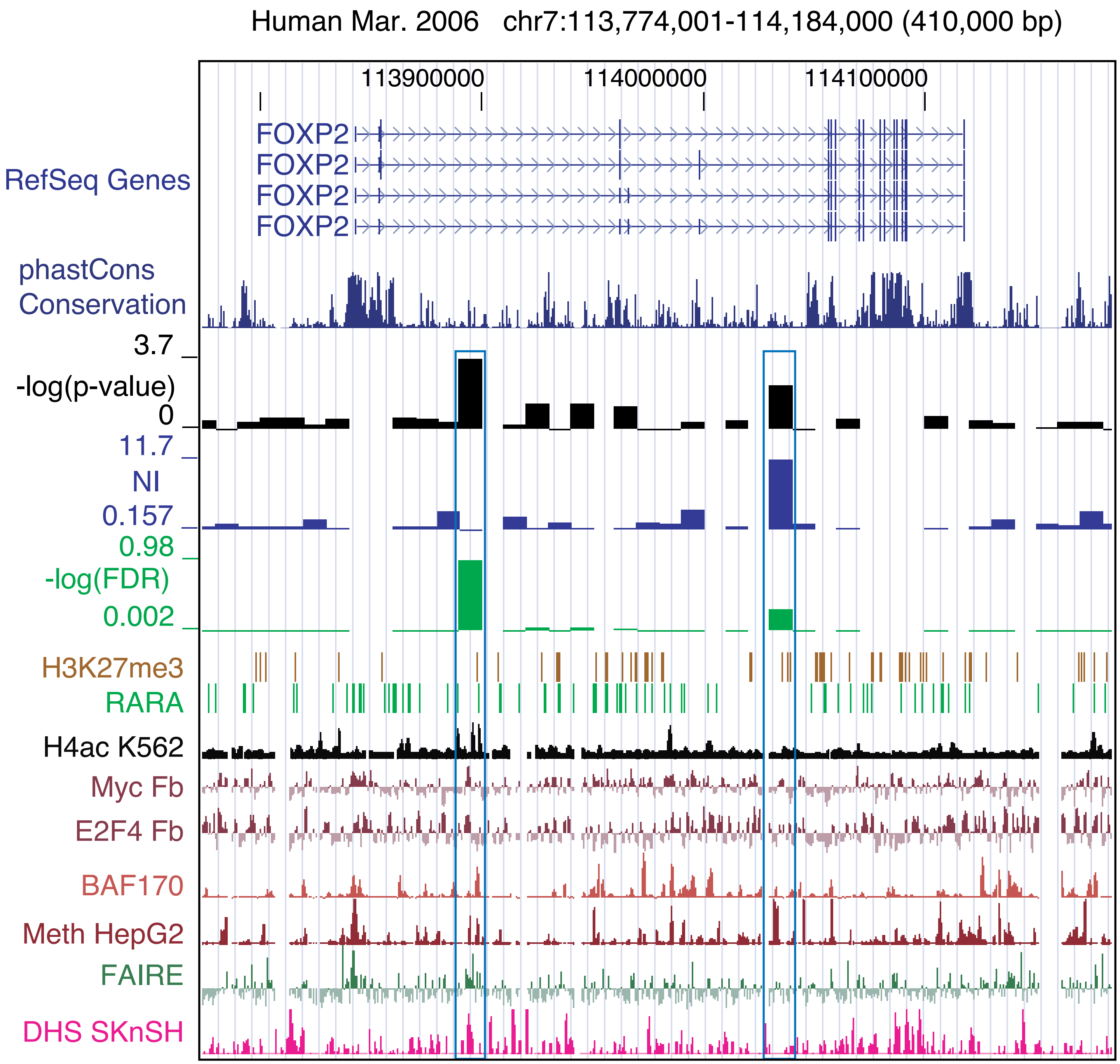
C. Statistics on the interval

	nonAR sites	AR sites
polymorphisms	15	13
divergence from rhesus	433	80
r_{pd}	0.035	0.16

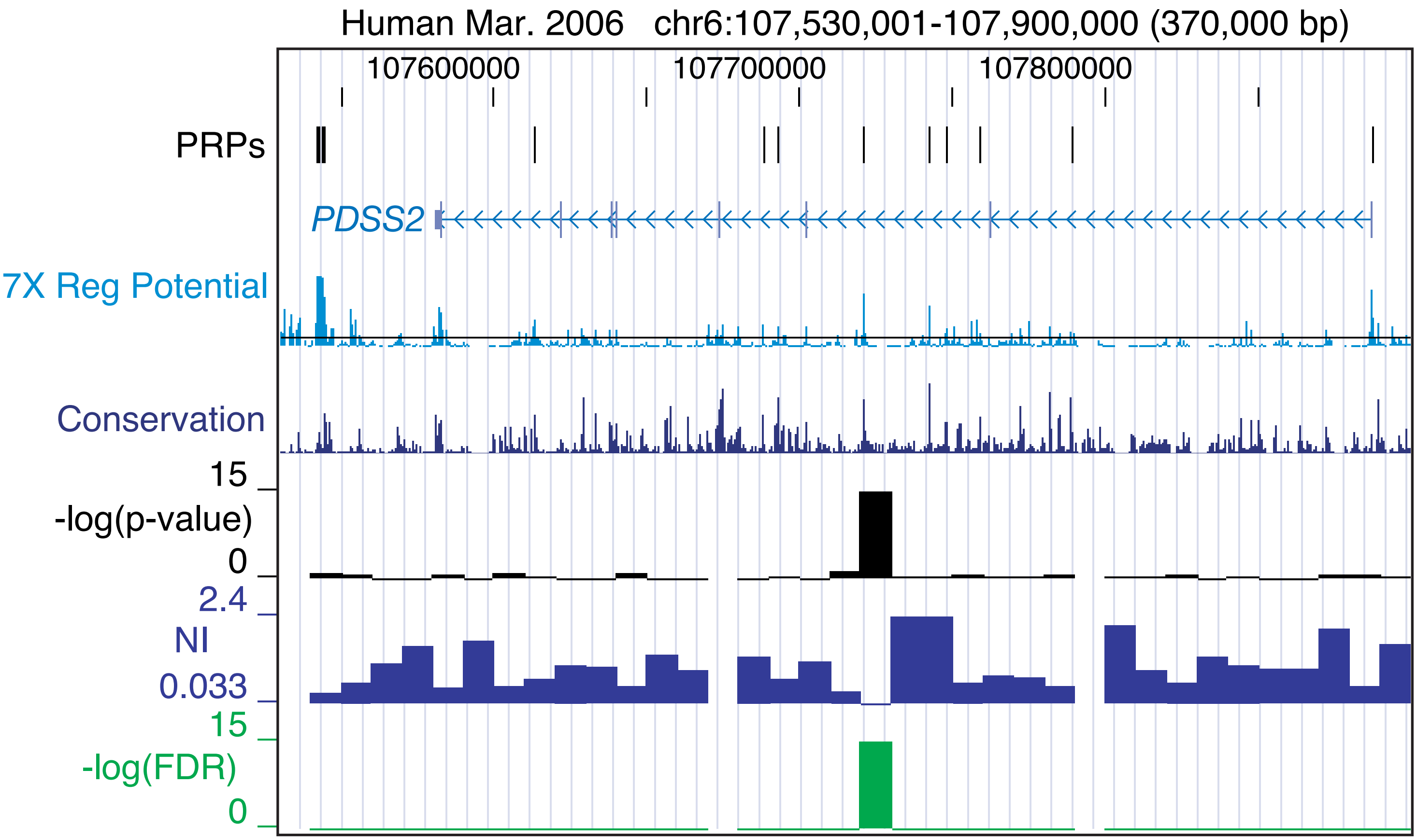
chi-square p-value = 0.000077

FDR = 0.02

NI=0.21



	nonAR sites	AR sites	nonAR sites	AR sites
polymorphisms	8	10	13	1
divergence from chimp	142	27	41	37
r_{pd}	0.056	0.37	0.31	0.027
p-value	0.00022		0.0063	



	nonAR sites	AR sites
polymorphisms	26	67
divergence from rhesus	493	42

r_{pd} 0.052 1.59

p-value = 0.000000

FDR = 0.000