Divergent Evolution of Human p53 Binding Sites: Cell Cycle Versus Apoptosis

Monica M. Horvath, Xuting Wang, Michael A. Resnick, Douglas A. Bell*

Laboratory of Molecular Genetics, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, United States of America

The p53 tumor suppressor is a sequence-specific pleiotropic transcription factor that coordinates cellular responses to DNA damage and stress, initiating cell-cycle arrest or triggering apoptosis. Although the human p53 binding site sequence (or response element [RE]) is well characterized, some genes have consensus-poor REs that are nevertheless both necessary and sufficient for transactivation by p53. Identification of new functional gene regulatory elements under these conditions is problematic, and evolutionary conservation is often employed. We evaluated the comparative genomics approach for assessing evolutionary conservation of putative binding sites by examining conservation of 83 experimentally validated human p53 REs against mouse, rat, rabbit, and dog genomes and detected pronounced conservation differences among p53 REs and p53-regulated pathways. Bona fide NRF2 (nuclear factor [erythroid-derived 2]-like 2 nuclear factor) and NFκB (nuclear factor of kappa light chain gene enhancer in B cells) binding sites, which direct oxidative stress and innate immunity responses, were used as controls, and both exhibited high interspecific conservation. Surprisingly, the average p53 RE was not significantly more conserved than background genomic sequence, and p53 REs in apoptosis genes as a group showed very little conservation. The common bioinformatics practice of filtering RE predictions by 80% rodent sequence identity would not only give a false positive rate of ~19%, but miss up to 57% of true p53 REs. Examination of interspecific DNA base substitutions as a function of position in the p53 consensus sequence reveals an unexpected excess of diversity in apoptosis-regulating REs versus cell-cycle controlling REs (rodent comparisons: $p < 1.0$ e–12). While some p53 REs show relatively high levels of conservation, REs in many genes such as *BAX, FAS, PCNA, CASP6, SIVA1*, and *P53AIP1* show little if any homology to rodent sequences. This difference suggests that among mammalian species, evolutionary conservation differs among p53 REs, with some having ancient ancestry and others of more recent origin. Overall our results reveal divergent evolutionary pressure among the binding targets of p53 and emphasize that comparative genomics methods must be used judiciously and tailored to the evolutionary history of the targeted functional regulatory regions.

Citation: Horvath MM, Wang X, Resnick MA, Bell DA (2007) Divergent evolution of human p53 binding sites: Cell cycle versus apoptosis. PLoS Genet 3(7): e127. doi:10.1371/journal.pgen.0030127

Introduction

Since the completion of the human genome, cataloging transcription factor binding sites (TFBSs) has been critical for understanding gene regulation. The use of comparative genomics (evolutionary conservation across species) is often championed as a method to separate the functional regulatory sequence "wheat" from the nonfunctional "chaff" [1]. As the number of mammalian full genome drafts increases, the integration of TFBS predictions with lists of conserved noncoding regions (CNCs) has emerged as a key step in the TFBS identification process [2–4]. If TFBS predictions are contiguous to DNA features that coincidentally have a critical structural role such as maintenance of chromatin organization, the appearance of conservation may be intensified even further.

Although these methods have greatly enhanced our knowledge of the human genome's regulatory repertoire, overreliance on conservation information can potentially exclude genuine binding sites. Since TFBSs are typically small, they can arise by chance in a gene's promoter and therefore may decrease selective pressures to maintain already existing sites [5]. Another wrinkle is that the evolutionary forces that created the conservation blocks may no longer be functionally relevant to humans [6]. Additionally, recent scans for natural selection in human gene coding regions have revealed that distinct biological pathways often are subject to widely different evolutionary pressures [7,8], particularly since mutation rates have been shown to vary across the genome [9]. Genes involved in oncogenesis and tumor suppression have experienced recent selection for mutation in primate lineages [7,8,10]. DNA binding sites of transcription factors

Editor: Takashi Gojobori, National Institute of Genetics, Japan

Received January 31, 2007; **Accepted** June 15, 2007; **Published** July 27, 2007

A previous version of this article appeared as an Early Online Release on June 15, 2007 (doi:10.1371/journal.pgen.0030127.eor).

Abbreviations: AUC, area under the curve; CNC, conserved noncoding region; NFκB, nuclear factor of kappa light chain gene enhancer in B cells; NRF2; nuclear factor (erythroid-derived 2)-like 2 nuclear factor; OR, odds ratio; RE, response element; ROC, receiver operator characteristic; TFBS, transcription factor binding site

* To whom correspondence should be addressed. E-mail: BELL1@niehs.nih.gov

Author Summary

The p53 tumor suppressor is a transcription factor that coordinates cellular responses to DNA damage and stress, initiating cell-cycle arrest or triggering apoptosis. Evolutionary conservation is often employed to separate the functional "wheat" from the nonfunctional "chaff" when identifying binding sites of transcription factors like p53. We evaluated evolutionary conservation of 83 experimentally validated human p53 binding sites against mouse, rat, rabbit, and dog genomes, and similarly examined binding sites for two other transcription factors as controls, NRF2 (nuclear factor [erythroid-derived 2]-like 2 nuclear factor) and NFκB (nuclear factor of kappa light chain gene enhancer in B cells), which direct oxidative stress and innate immunity responses, respectively. NRF2 and NFκB binding sites both exhibited high interspecific conservation, indicative of purifying natural selection, but surprisingly, human p53 response elements on average displayed a lack of conservation. Thus conservation is not useful in the prediction of p53 binding sites. After grouping p53 REs by gene ontology, we observed that binding sites in cell-cycle genes like *CDKN1A* displayed high conservation, while p53 binding sites in apoptosis and DNA repair genes showed an unexpected excess of diversity and very little homology with rodent sequences. Overall these results reveal divergent evolutionary pressure among the binding targets of p53 and suggest caution in generalizing about the similarity of regulation of the p53 pathway between humans and rodents.

are also functional components of these pathways and are likely under similar evolutionary pressures. Indeed, we have focused recently on identifying human single nucleotide polymorphisms that alter the function of transcription factors [11,12]. As a result, we have investigated the assumptions for using mammalian conservation as an obligatory screening step for seeking TFBSs.

The p53 tumor suppressor gene, encoded by the *TP53* master regulatory gene, is a transcription factor that coordinates a network of cellular responses to environmental insults. Over half of human cancers have a mutation in the p53 protein or one of its partners [13]. The p53 protein is estimated to have several hundred transregulation target genes that affect pathways including apoptosis, DNA damage repair, and cell-growth arrest [14]. As a result, p53 target genes are highly sought-after drug targets for halting cancer progression. According to in vitro experiments, the p53 protein binds specifically to a palindromic consensus sequence, RRRCWWGYYY(N_{0-13})RRRCWWGYYY [15], with nearly all REs containing at least one mismatch; in vivo results have suggested that the spacer region may be much smaller [14,15]. The sequence is typically located within 5,000 bases of the target gene's transcriptional start site, and p53 either induces or represses expression upon p53 binding [16,17]. One feature of p53 that confounds the discovery of novel transregulated genes is that while some binding sites match the expected consensus sequence quite well, others can be consensus poor and yet are both necessary, and sufficient, to transactivate a gene [18]. A recent study has suggested that the "rules of engagement" for p53 REs may differ based on the activated pathway, particularly in the apoptosis and cell-cycle–related systems [19]. Thus, we have used cross-species conservation to examine if these groups of elements exhibit distinct conservation profiles.

To evaluate the utility of comparative genomics approaches in the identification of potential p53 target REs, we gleaned the literature for a high quality set of bona fide p53 REs to estimate the degree of conservation between humans and other mammals. To relate the *TP53* system to other master regulators, we compare its binding site conservation to those of the transcription factors encoded by two other genes: *NFκB* (nuclear factor of kappa light chain gene enhancer in B-cells), central to inflammation responses, and *NFE2L2,* which encodes NRF2 (nuclear factor [erythroid-derived 2]-like 2 nuclear factor), a regulator of oxidative stress. Their repertoire of interactions is expected to be highly preserved throughout the mammalian lineage. The NFκB transcription factor is a heavily studied biological switch of the inflammation, apoptosis, and immune responses [20,21]. It binds the consensus sequence GGGRNNYYCC [22,23], and its signaling system is highly conserved even when examined in invertebrates [21,24]. NRF2 binds antioxidant REs (consensus sequence = TMANNRTGAYNNNGCRWWWW [25]) that are comparable in size to those of p53, show high levels of conservation [26], and are found in the promoters of genes that confer protection from oxidative stress and chemical carcinogens [27]. Mouse models of Nrf2-dependent response to oxidative and electrophilic insults have been used to study function [28,29]. Additionally, the Nrf2 pathway in zebrafish operates similarly to humans and underscores the likelihood of high conservation in regulatory binding sites [30]. Because the NFκB and NRF2 binding sites were determined to be highly conserved, these two sets of TFBSs serve as positive controls in estimates of conservation. Our comparative genome analysis, which includes a coincident evaluation of sampled promoter sequences and coding region sequence, reveals that mammalian conservation does not apply to p53 target REs in general. However, among subgroups of target genes we observe purifying selection acting on a number of p53 binding sites, including many cell-cycle–related genes, while rodent to human homology is lacking for p53 REs in apoptosis-related genes.

Results

Conservation of TFBSs

The literature was scanned for validated p53 (83, Table 1), NFκB (21, Table 2), and NRF2 (21, Table 3) binding sites associated with human genes. Human genome coordinates were located and then referenced within global multiple alignments held at the University of California Santa Cruz [UCSC] (California, United States) genome browser website [31] to find their corresponding locations in eleven other mammals. Using these global alignments, percent sequence identity was calculated for each of the 125 binding sites across the eleven mammals, with the calculation adapted to reflect consensus sequence degeneracy, since every model RE had positions where one or more of the four nucleotides could be tolerated. Also, the p53 RE was unique in that the spacer region between the two half sites could be any size or sequence up to thirteen bases. We therefore calculated sequence identity by omitting the p53 spacer region and ignoring mismatches in the alignments that still fit the TFBS consensus (*CDKN1A* and *PCNA* examples shown in Figure 1, others in Figure S3).

Figure 2 plots the conservation distribution for each set of human TFBSs across mouse, rabbit, rat, and dog. Although

Table 1. Validated Human p53 TFBSs

Target Gene	Sequence	Function	Target Gene	Sequence	Function
APAF1	AGACATGTCT(13)CGACAAGCCC	Apoptosis	*BTG2*	AGTCCGGGCA(1)AGCCCGAGCA	Cell cycle
BAX	TCACAAGTTA(1)AGACAAGCCT	Apoptosis	*CDKN1A*	AAACATGCCC(11)CAACAAGCTG	Cell cycle
BAX	AGACAAGCCTGGGCGTGGGC	Apoptosis	*CDKN1A*	GAAGAAGACTGGGCATGTCT	Cell cycle
BBC3	CTGCAAGTCCTGACTTGTCC	Apoptosis	*CDKN1A*	GAACATGTCCCAACATGTTG	Cell cycle
BID	GGGCATGATGGTGCATGCCT	Apoptosis	*DUSP1*	GAACTTGTCA(2)GGCTTTGTTT	Cell cycle
CASP1	AGACATGCATATGCATGCAC	Apoptosis	*GDF15*	CATCTTGCCCAGACTTGTCT	Cell cycle
CASP6	AGGCAAGGAG(4)AGACAAGTCT	Apoptosis	*GDF15*	AGCCATGCCCGGGCAAGAAC	Cell cycle
CTSD	AAGCTGGGCCGGGCTGACCC	Apoptosis	*MDM2*	GGTCAAGTTCAGACACGTTC	Cell cycle
FOS	GCGTGCGCTC(1)GAGCAAGTCC	Apoptosis	*MDM2*	GAGTTAAGTC(1)TGACTTGTCT	Cell cycle
FOS	ACGCTTGCCATAGTAAGAAT	Apoptosis	*PLK2*	AGACATGGTG(3)AAACTAGCTT	Cell cycle
DCC1	CAGCATGTTCACACAAGCCA	Apoptosis	*PLK2*	AAACATGCCTGGACTTGCCC	Cell cycle
FAS	GGACAAGCCCTGACAAGCCA	Apoptosis	*PLK2*	GGTCATGATT(3)TAACTTGCCT	Cell cycle
IGFBP3	AAACAAGCCA(1)CAACATGCTT	Apoptosis	*PLK3*	TAACATGCCC(6)AAGCGAGCGC	Cell cycle
IGFBP3	GGGCAAGACCTGCCAAGCCT	Apoptosis	*PTEN*	GAGCAAGCCC(14)GGGCATGCTC	Cell cycle
LRDD	AGGCCTGCCT(8)GGACATGTCT	Apoptosis	*SESN1*	GGACAAGTCTCCACAAGTCT	Cell cycle
P53AIP1	TCTCTTGCCCGGGCTTGTCG	Apoptosis	*SFN*	TAGCATTAGC(2)AGACATGTCC	Cell cycle
PCBP4	GAACTTAAGA(10)GGACAAGTTG	Apoptosis	*TP73*	GGGCAAGCTGAGGCCTGCCC	Cell cycle
PERP	AGGCAAGCTCCAGCTTGTTC	Apoptosis	*DKK1*	AGCCAAGCTT(6)AACCAAGTTC	Cell growth
PLAGL1	CAACTAGAC(4)TAGACTAGCTT	Apoptosis	*EDN2*	CTGCAAGCCCGGGCATGCCC	Cell growth
PMAIP1	GAGCGTGTCCGGGCAGGTCG	Apoptosis	*FLT1*	GGACACGCTC(5)GGACCTGAGC	Cell growth
TNFRSF10B	GGGCATGTCCGGGCAAGACG	Apoptosis	*NDRG1*	CCACATGCAC(12)GCACATGAAC	Cell growth
TNFRSF10C	GGGCATGTCCGGGCAGGACG	apoptosis	*SERTAD1*	GGGCATGCGCCCTGAAGCCC	Cell growth
TP53I3	CAGCTTGCCCACCCATGCTC	Apoptosis	*TGFA*	GGGCAGGCCCTGCCTAGTCT	Cell growth
TP53INP1	GAACTTGGGGGAACATGTTT	Apoptosis	*DDB2*	AAGCTGGTTTGAACAAGCCC	DNA repair
TRAF4	GGGCAAGCCAGGGCCTGCCT	Apoptosis	*GADD45A*	GAACATGTCTAAGCATGCTG	DNA repair
SIVA	CTACATGCACACGCATGCAT	Apoptosis	*MLH1*	AGGCATGTAC(2)CGCATGCCCA	DNA repair
SIVA	GTACTTGGCAGGGCATGTCT	Apoptosis	*PCNA*	ACATATGCCCGGACTTGTTC	DNA repair
SIVA	GCACAAGCCT(5)TGTCTGGAGG	Apoptosis	*PMS2*	ATACTTGCTT(2)TTTCTTGTAA	DNA repair
WIG1	AAACAAGTCCAGACATGCCT	Apoptosis	*RRM2B*	TGACATGCCCAGGCATGTCT	DNA repair
ADARB1	GTGCAAGTTTCAACTTGTCC	RNA processing	*XRCC5*	GAACTAGTTT(1)AAACATGTTC	DNA repair
ARHGEF7	AAACATGTCAGCACTTGCTT	Cell signaling	*ATF3*	AGTCATGCCG(17)GGTCATGCCT	Transcription factor
DUSP5	GCACTAGACA(1)GGGCTTGTTG	Cell signaling	*C12ORF5*	AGACAAGTCT(2)GGACATGTCT	Metabolism
PDGFC	GGTCATGTTCAGACTTGCCC	Cell signaling	*PRKAB1*	GTTCTTGCCGCGGCTTGCCT	Metabolism
PPM1J	GAACATGCCTGAGCAAGCCC	Cell signaling	*SCGB1D2*	GGTCTTGTTTAGACTTGCTC	Unknown
DSC3	GAAGTTGCTC(6)AGGCAAGCCT	Cell adhesion	*PHLDA3*	GGTCAAGTTC(6)CAGCAGGCCA	Unknown
MASPIN1	GAACATGTTG(1)AGGCCTTTTG	Cell adhesion	*SERPINE1*	ACACATGCCTCAGCAAGTCC	Blood coagulation
MASPIN1	CAAGCTGCCA(2)AGGCTTGAGT	Cell adhesion	*DDIT4*	AAACAAGTCTTTCCTTGATC	ROS regulation
MMP2	AGACAAGCCTGAACTTGTCT	Angiogenesis	*FDXR*	GGGCAGGAGCGGGCTTGCCC	ROS regulation
SEMA3B	TTGCATGCCC(2)AGACATGTCT	Development	*GPX1*	GGGCCAGACCAGACATGCCT	ROS regulation
EOMES	GGGCCTGTCT(1)CAACTTGCCC	Morphogenesis	*SCARA3*	GGGCAAGCCCAGACAAGTTG	ROS regulation
UBTD1	GAGCAAGCCCAGACTTGTCA	Protein modification	*SOD2*	GTGCTTGTTC(4)GGGCATGTCC	ROS regulation
			RPS27L	GGGCATGTAGTGACTTGCCC	Translation

p53 consensus sequence = RRRCWWGYYY(N_{0-13})RRRCWWGYYY. The number of spacer bases are shown in parentheses. R = G or A; W = A or T; Y = C or T. ROS, Reactive oxygen species.

doi:10.1371/journal.pgen.0030127.t001

comparative data relating the chimpanzee and rhesus monkey were also analyzed, we observed, as expected, that these species were too evolutionary close to humans to be informative (Figure S1). For example, nearly any human sequence was in excess of 95% conserved in these two primates. Also graphed in Figure 2 are the results for sets of DNA sequence fragments randomly chosen from promoter (gray) and protein-coding (blue) regions. This allows TFBS conservation levels to be viewed in context of the evolutionary pressures exerted on other genomic sequences. The poorly conserved element shown in Figure 1 *(PCNA)* would fall in a lower percentage bin (i.e., be on the left side of the graph), as is the case for the randomly chosen promoter sequences. The promoter fragments are representative of the background genome sequence in which most TFBSs reside. On the other hand, if TFBSs were very well conserved, then the distribution would be right-shifted, as is the case for the protein-coding region fragments (blue). To use conservation as a metric to separate true binding sites from the rest of the genome, their conservation should be significantly greater than that of randomly chosen promoter regions. The spike of TFBSs at 0%–9% identity in each panel represents species-specific sites that are essentially not present in the other mammals.

For each of the human-to-mammal comparisons in Figure 2, *NFκB* and NRF2 sites produced identity distributions that appear very similar to distributions from the coding region fragment group (many sites with 90%–100% identity), which was representative of genome sequence under high purifying selection. This strongly suggests that *NFκB* and NRF2 sites may be under purifying selection. The human-to-mammal p53 site comparisons, on the other hand, produced conservation profiles in each species that have a high frequency of sites at zero percent identity and fewer with 90%–100%

Table 2. Validated Human NRF2 TFBSs

Target Gene	Sequence	Target Gene	Sequence
ABCC1	TCTGTGTGACTCAGCTTTGGA	*HMBS*	CTCCAGTGACTCAGCACAGGT
ETS1	AGCGGGTGACCAAGCCCTCAA	*NQO1*	TCACAGTGACTCAGCAGAATC
FTH1	CCTCCATGACAAAGCACTTTT	*S100A6*	GACACGTGACTCGGCAAGGGG
FTL	TCAGCATGACTCAGCAGTCGC	*SAT*	CCGCTATGACTAAGCGCTAGT
GCLC	TCCCCGTGACTCAGCGCTTTG	*SPTA1*	ACTGGGTGACTCAGCAGTTTT
GCLM	TAACGGTTACGAAGCACTTTC	*TBXAS1*	AAGGAATGAATCAGCAACTTT
GCLM	AGACAATGACTAAGCAGAAAT	*TXNRD1*	TCAGAATGACAAAGCAGAAAT
GNAI2	AGCCTGTGACTGGGCCGGGGC	*TXNRD1*	TCATTCTGACTCTGGCAGTTA
GPX2	CCAGGATGACTTAGCAAAAAC	*UGT1A6*	TCTGTCTGACTTGGCAAAAAT
GSTP1	GCGCCGTGACTCAGCACTGGG	*UGT1A6*	GAAAGCTGACACGGCCATAGT
HBB	TCATCATGACTCAGCATTGCT		

NRF2 consensus sequence = TMAnnRTGAYnnnGCRWWW.
doi:10.1371/journal.pgen.0030127.t002

identity. This distribution is similar to the distribution obtained from randomly sampled promoter fragments (gray), which we used to represent genome sequence under neutral selection. In mouse the p53 RE identity distribution was correlated with the promoter fragment identity distribution while NFκB and NRF2 showed less correlation with the promoter distribution (Table S3).

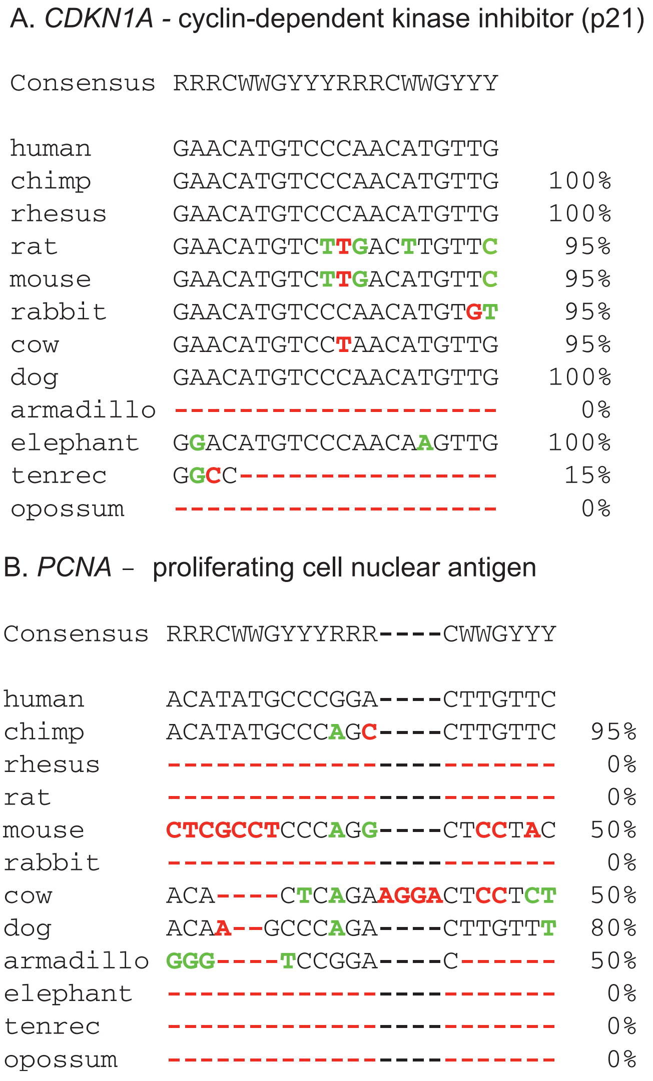
Since p53 sites as a group were observed to have as many interspecific substitutions as the background genomic sequence, use of conservation level to predicting bona fide sites would not be effective. However, this result could be due to the fact that the set of 83 p53 TFBSs actually represents two or more subsets of p53 REs with distinct conservation profiles. A recent study hypothesized that the sequence requirements of p53 REs may differ based on the activated pathway such as apoptosis, DNA repair, cell-cycle checkpoints, or cell-growth arrest [19]. We therefore investigated if low and high percent identity values would apportion with p53 REs grouped by function, thereby detecting evolutionary divergence between p53 pathways.

Among these 83 p53 REs, we carried out analysis of the two largest subgroups (Table 1), apoptosis-related ($n = 29$) and cell-cycle/cell-growth–related ($n = 23$), on the basis of observations of Qian et. al [19]. Average percent identity to a consensus sequence was calculated for each of the transcription factor groups, including the p53 subsets, and compared via a two-tailed t-test assuming unequal variances between the datasets. The results are displayed as odds ratios (OR) in Table 4, and OR values represent the odds that one type of human TFBS (columns) will be found as more conserved than a second TFBS type (rows) in comparison with other mammalian species. Ratios less than 1 (e.g., p53 apoptosis compared to all p53 sites) suggest lower conservation of the TFBS in the row. Among all species, the relative conservation levels of NFκB compared to NRF2 sites were similar and the p-values for difference were not significant. NFκB sites were significantly more conserved than the entire set of p53 sites in mouse and dog, while NRF2 sites were significantly more conserved than all p53 sites in mouse and rat. The magnitude and statistical significance of the differences between sequence motifs in Table 4 was greatest when comparing either NRF2 or NFκB and the apoptotic p53 sites. For example, in each species, the ORs for relative conservation between either NRF2 or NFκB and apoptosis genes were all high (OR > 3.0) and at a significance of at least $p < 0.019$. On the other hand, the mean conservation level of the cell-cycle–regulating p53 REs were not statistically different from the NRF2 or NFκB sites. These observations imply that the

Table 3. Validated Human NFκB TFBSs

Target Gene	Sequence	Target Gene	Sequence
SELE	GGGGATTTCC	*TRAF1*	GGGGAACTCT
PTGS2	GGGGATTCCC	*TNIP1*	GGAAAGTCCC
MAT2A	GGGAGGTGCC	*NFKBIA*	GGAAATTCCC
CCL20	GGGGAAAACC	*ALOX12*	GGGACATCCC
IL8	TGGAATTTCC	*TP53*	GGGGTTTTCC
CXCL1	GGGAATTTCC	*ICAM1*	GGAAATTCCG
IRF1	GGGAATCCCG	*ZFP36*	GGGGATCCGA
LTA	GGGGCTTCCC	*SDC4*	GGGGAATTCC
IL6	TGGGATTTTC	*ELK1*	GGGAAGGGCC
SERPINE1	TGGAATTTCT	*F8*	GGGACTTTTC
TNFRSF10B	GGGAATTCCC		

NFκB consensus sequence = GGGRnnYYCC.
doi:10.1371/journal.pgen.0030127.t003



A. *CDKN1A* - cyclin-dependent kinase inhibitor (p21)

```
Consensus  RRRCWWGYYYRRRCWWGYYY

human      GAACATGTCCCAACATGTTG
chimp      GAACATGTCCCAACATGTTG   100%
rhesus     GAACATGTCCCAACATGTTG   100%
rat        GAACATGTCTTGACTTGTTC    95%
mouse      GAACATGTCTTGACATGTTC    95%
rabbit     GAACATGTCCCAACATGTGT    95%
cow        GAACATGTCCTAACATGTTG    95%
dog        GAACATGTCCCAACATGTTG   100%
armadillo  --------------------     0%
elephant   GGACATGTCCCAACAAGTTG   100%
tenrec     GGCC----------------    15%
opossum    --------------------     0%
```

B. *PCNA* – proliferating cell nuclear antigen

```
Consensus  RRRCWWGYYYRRR----CWWGYYY

human      ACATATGCCCGGA----CTTGTTC
chimp      ACATATGCCCAGC----CTTGTTC   95%
rhesus     ------------------------    0%
rat        ------------------------    0%
mouse      CTCGCCTCCCAGG----CTCCTAC   50%
rabbit     ------------------------    0%
cow        ACA----CTCAGAAGGACTCCTCT   50%
dog        ACAA--GCCCAGA----CTTGTTT   80%
armadillo  GGG----TCCGGA----C------   50%
elephant   ------------------------    0%
tenrec     ------------------------    0%
opossum    ------------------------    0%
```

Figure 1. Examples of Calculated Interspecies Conservation Scores for TFBSs

Multispecies alignments for p53 binding sites in the (A) *CDKN1A* and (B) *PCNA* genes are shown. Mismatches (green) from human that do not alter the p53 consensus motif, RRRCWWGYYY($N_{0–13}$)RRRCWWGYYY, do not penalize the percent sequence identity, whereas consensus-altering mismatches and insertion/deletion events do (red). For the p53 binding sites, spacer elements are in gray and are not considered when calculating percent identity. R = G or A; W = A or T; Y = C or T.

doi:10.1371/journal.pgen.0030127.g001

mean p53 conservation level for all elements is really a combination of the effect of the two ontological subgroups.

It has been proposed that an alternative method for evaluating a DNA fragment's conservation level is to ask whether it sits within a block of conservation [6]. All bona fide TFBSs examined in this study were matched against the "most conserved" track of the UCSC genome browser to ask whether a significant proportion fell within conserved blocks. Only 19.2% of all p53 REs mapped to these regions, while 52.8% of NRF2 and 57.1% of NFκB TFBSs could be colocated (Table 5). We also assessed the conservation of the randomly chosen promoter sequences according to this block method and used a two-tailed binomial test to calculate statistical significance. Intriguingly, all TFBS groups mapped to more blocks than the random promoter sequences except for the apoptosis subgroup of p53 REs. These data mirrored the percent identity conservation metric and again underscored that the apoptotic-regulating p53 TFBSs may not have been under purifying selection throughout mammalian evolution.

Inspection of the individual alignments between human p53 REs and mouse reveal that 38% (11/29) apoptosis elements and 9% (2/21) of cell-cycle element could not be aligned with the multiz global alignment tool and thus showed zero identity. Thus for human p53 REs in genes such as *BAX, FAS, PCNA, CASP6, SIVA1,* and *P53AIP1* we observed little, if any, similarity with rodent sequences, and these nonaligned sequences (zero identity) strongly impact the calculations we have made.

Predictiveness of Mammalian Conservation for TFBSs

Although it was informative to know how well human p53 TFBSs are conserved relative to other regulatory motifs, an aim of this study was to probe the utility of comparative genomics for authenticating binding sites predicted by computational methods. Receiver operator characteristic (ROC) curves were employed to demonstrate the sensitivity of TFBS prediction when qualified by conservation information. ROC curves are traditionally used to measure the quality of a binary classification algorithm, as a discrimination threshold is varied. Area under a ROC curve provides a visual representation of how well the conservation metric can classify the sets of bona fide TFBSs as true positives. For example, the area can be interpreted as the probability that when both a bona fide TFBS and a random promoter sequence of equal lengths are chosen at random, the decision function (conservation in a species) assigns a higher value to the bona fide TFBS. A perfect decision function would generate a curve with an area of 1, meaning that 100% sensitivity was obtained (i.e., all true positives were found), and 100% specificity was reached (i.e., no false positives were generated). If conservation predicted TFBS authenticity no better than random chance, a line at 45 ° to the x-axis would be generated that bisects the ROC space (area under the curve [AUC] = 0.5), because as the threshold is raised, equal numbers of true and false positives compose the chosen set of TFBSs. A ROC curve that fell below this diagonal would indicate that conservation consistently predicted poorly, meaning that one should employ the lack of conservation as a decision classifier to authenticate TFBSs.

When TFBS conservation was evaluated as a TFBS predictor in each of the four mammals, bona fide NRF2 and NFκB sites were consistently well predicted, whereas the ROC curve describing all p53 sites approached the random diagonal (Figure 3A–3D). The latter implies that conservation analysis in these model organisms cannot enhance p53 binding site discovery, for the predictive capacity is only slightly better than random. For example, if a cutoff of 80% identity to mouse was employed as the rule for choosing p53 binding sites, only 43% of real p53 REs would be found, and 19% of the selections would be false positives (Figure 3A). We were concerned that ascertainment bias (e.g., the presence of spurious REs) in this large set of p53 sites might affect our findings. However, the predictivity level for any of these species does not change appreciably even when the p53 RE list is restricted to only the 30 best-characterized sites (Figure

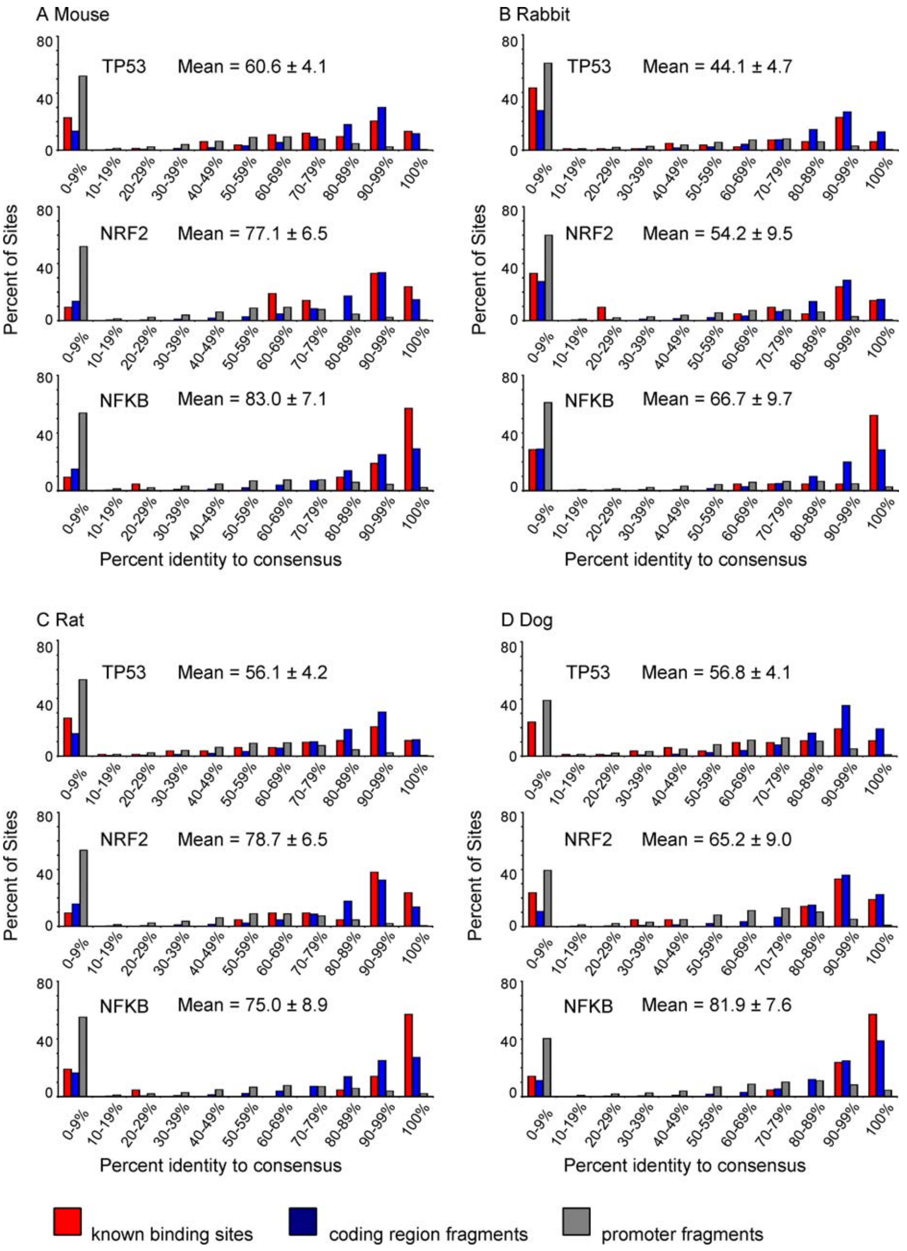


Figure 2. Distribution of Human-to-Mammal TFBS Conservation Scores

Graphed for each (human-to-mouse, -rabbit, -rat, and -dog) comparison in (A–D) are: the percentage of all TFBSs from each gene group (p53, 83 REs; NRF2, 21 REs; or NFκB, 21 REs) that fall within the x-axis conservation bins (red bars); data for randomly sampled fragments from human coding regions (blue bars); and gene promoters (gray bars). Coding regions (representing sequence under purifying selection) and randomly sampled promoter regions (representing sequence under more neutral selection) were sampled 1,000 trials; error bars are narrower than the line art. Averages and standard errors for the sets of known TFBSs are reported above each graph.

doi:10.1371/journal.pgen.0030127.g002

S2). In contrast, for NFκB an 80% mouse conservation threshold would allow discovery of 86% of real NFκB sites with a 25% false positive rate (Figure 3A). As a result, human/mouse multiple sequence alignments are highly useful for identifying novel NFκB sites but not so for p53 motifs. This conclusion is reiterated in all four model organisms by statistical evaluation of the AUC calculations (Table 6). p53 curves had smaller AUCs compared to those for NFκB and NRF2.

A recent study noted that although the spacer region between half sites for p53 REs can be zero to13 bases, small spacers were overwhelmingly preferred in a distribution of spacer length derived from genome-wide chromatin immunoprecipitation experiments [14]. This suggests that REs with large spacers might not be valid, and we hypothesized that if ROC curve analysis was restricted to only p53 REs with small spacers (presumed higher quality), perhaps much greater conservation would be observed. We examined the set of p53 REs having two or fewer spacer nucleotides between half sites and observed no increase in conservation (Figure S2). Not only was there no improvement in TFBS prediction sensitivity for this subset, but the ROC areas were visibly greater

Table 4. Comparison of Mean TFBS Conservation Values (Mouse to Human, Rabbit to Human, Rat to Human, and Dog to Human)

Animal	TFBS	OR[a] (*p*-value*)			
		TP53 (Apoptosis)	TP53 (Cell Cycle)	NRF2	NFKB
Mouse	TP53 (all)	0.63 (0.218)	1.74 (0.101)	2.20 (0.038)*	3.18 (0.009)*
	TP53 (apoptosis)		2.73 (0.023)*	3.44 (0.009)*	4.99 (0.003)*
	TP53 (cell cycle)			1.26 (0.626)	1.82 (0.277)
	NRF2				1.45 (0.544)
Rabbit	TP53 (all)	0.38 (0.006)*	1.60 (0.261)	1.39 (0.438)	2.54 (0.393)
	TP53 (apoptosis)		4.21 (0.006)*	3.67 (0.015)*	6.68 (<0.001)*
	TP53 (cell cycle)			0.67 (0.793)	1.58 (0.421)
	NRF2				1.82 (0.548)
Rat	TP53 (all)	0.54 (0.105)	2.21 (0.020)*	2.89 (0.006)*	2.34 (0.066)
	TP53 (apoptosis)		3.77 (0.003)*	5.26 (<0.001)*	4.25 (0.007)*
	TP53 (cell cycle)			1.31 (0.586)	1.06 (0.919)
	NRF2				0.81 (0.743)
Dog	TP53 (all)	0.44 (0.020)*	1.33 (0.427)	1.40 (0.432)	3.37 (0.008)*
	TP53 (apoptosis)		3.05 (0.012)*	3.18 (0.019)*	7.69 (<0.001)*
	TP53 (cell cycle)			1.04 (0.933)	2.52 (0.107)
	NRF2				2.41 (0.165)

[a] The ORs are calculated as columns over rows. For example, in the human to mouse comparison, p53 cell cycle REs are 2.73 times more conserved than p53 apoptosis REs ($p = 0.023$).
*Significant *p*-values for a two-tailed t-test between each sample's mean conservation.
doi:10.1371/journal.pgen.0030127.t004

for rat and rabbit comparisons in the inverse set of REs (i.e., 3+ spacer bases). Thus, as judged by conservation, p53 RE spacer region length could not be considered a measure of RE quality.

We then examined p53 RE conservation in light of gene ontology. When p53 REs were subdivided based on functional class, the sensitivity of interspecific conservation to predict cell-cycle/cell-growth sites improved considerably, approaching that for the NFκB and NRF2 targeted genes (Figure 3). The p53 apoptotic REs (Figure 3, dashed line), on the other hand, showed a dramatically different conservation profile. In the case of the mouse (AUC = 0.616) and rat (AUC = 0.568), the ROC curve hovered just above the random line, which indicates a lack of sensitivity. For two species (rabbit, AUC = 0.469 and dog, AUC = 0.437), the sequence identity metric had an apoptosis RE discovery rate worse than random prediction. This suggests that the functional, apoptotic p53 binding sites are less conserved than randomly sampled sequences in gene promoters. This phenomenon was also observed when ROC curve analysis was carried out in other distant mammalian species (tenrec, armadillo, elephant, and opossum) (unpublished data).

Table 5. Mapping of Human TFBSs to Hidden Markov Model-Generated Conservation Blocks

TFBS	Number Conserved Binding Sites out of Total (%)	*p*-Value*
p53	16/83 (19.2%)	4.18e−5
p53, apoptosis	4/29 (13.8%)	0.22
p53, cell cycle	6/23 (26.1%)	0.002
NRF2	11/21 (52.8%)	4.00e−8
NFκB	12/21 (57.1%)	1.00e−9

*In a set of randomly chosen promoter regions of the same length, two-tailed binomial probability of finding an equivalent or greater fraction of conserved sites.
doi:10.1371/journal.pgen.0030127.t005

One explanation for this provocative result could be that apoptotic p53 sites might actually display a slightly different consensus p53 binding site than that reported in the literature. Perhaps a better-fitting consensus would improve conservation. We aligned all p53 sites (83), apoptotic (29), and cell-cycle (23) p53 sites (Figure S3) and generated sequence logos [32] (Figure S3A–3C) to identify improved patterns, but while there are small differences, none fit better than the existing consensus of RRRCWWGYYYN_{0-13}RRRCWWGYY. Likewise, simply permitting any nucleotide at the least compositionally biased positions in this p53 RE subset (bases 2, 8, 10, and 11 of the p53 consensus) did not improve the area under the ROC curve (Figure S2). These data emphasize not only that conservation analysis cannot improve identification of certain TFBSs like p53, but also that subclasses of the same binding site may reflect distinct evolutionary profiles.

Sequence Diversity as a Function of Position in the Consensus Sequence

A second approach was used to detect if conservation differed among nucleotide positions within the binding site. That is, could we observe heightened human-to-mammal interspecific substitutions or "sequence diversity" at particular locations within each TFBS consensus sequence? To accomplish this, we aligned all TFBSs within each group (p53, NRF2, or NFκB) and calculated the positional sequence diversity, which was the percentage of aligned bases at each position that varied from the accepted consensus sequence (Figure 4). For example, in Figure 4A, the first position in the p53 consensus sequence differed from a purine base (R) at the equivalent position in the mouse in 35% of all p53 TFBSs, while positional diversity for randomly sampled promoter sequence was 63% and that for coding region sequence was 25%. Highly conserved sequence would be plotted lower on

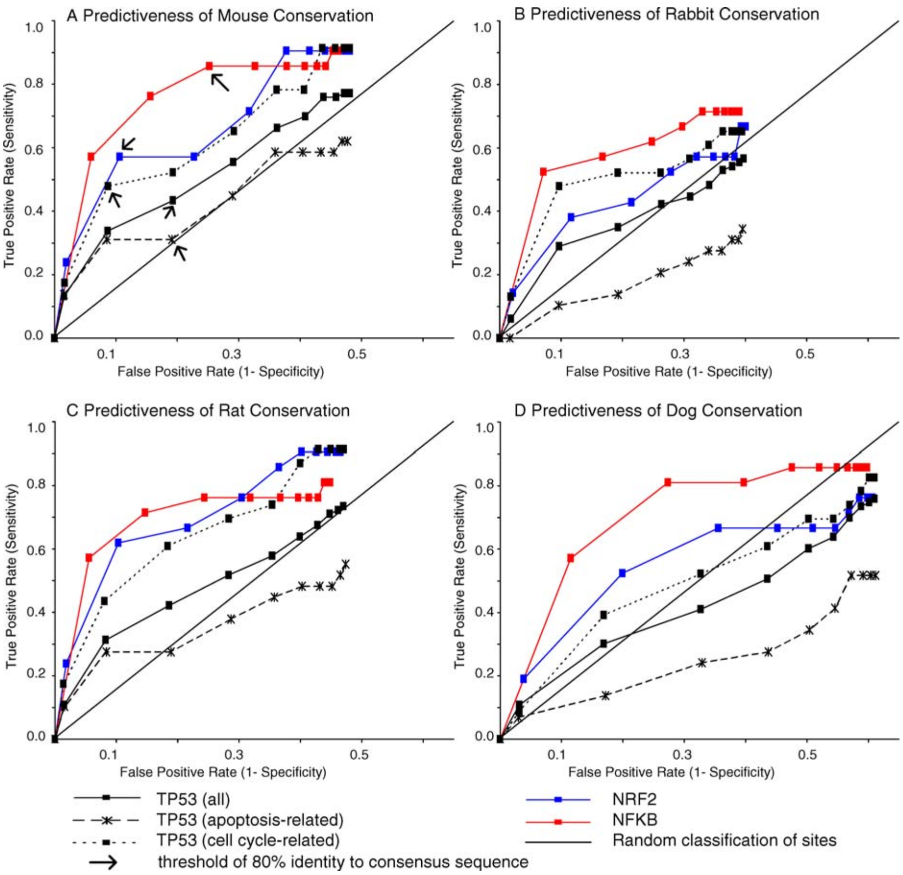


Figure 3. ROC Curves for Four Mammalian Comparisons Indicate the Ability of Different Interspecific Comparisons to Identify Bona Fide Human TFBSs

(A–D) Using a dataset of experimentally verified human TFBSs and randomly chosen promoter sequences from human genes, the true positive and false positive prediction rates were calculated given conservation thresholds that classify a TFBS as authentic or spurious. Each data point on a curve represents a different conservation threshold for comparison in the indicated species. In (A) arrows indicate the 80% identity threshold. For NFκB 80% identity with mouse is highly predictive for human. Prediction of apoptosis sites are similar to random classification (diagonal line).

doi:10.1371/journal.pgen.0030127.g003

the y-axis (less diversity) as displayed by the coding region line (blue), while less conserved sequence would exhibit high diversity and appear near the top of the graph (e.g., gray, random promoter sequences). Small peaks observed in the promoter and coding region plots reflect degeneracy of the consensus sequence, with more degenerate positions exhibiting less calculated diversity. Patterns of the promoter and coding sequence lines are highly similar across species in Figure 4 except for being shifted on the y-axis. This was an expected feature of the data since these control curves were plotted as the average result of 1,000 trials of sequence fragment sampling across the human genome. When examining the population statistic of a large number of fragments, the average coding or promoter region fragments will exhibit similar transversion and transition mutation rates across species, which are visualized in these patterns.

Figure 4D and 4E demonstrate that the positional sequence diversity of NRF2 and NFκB sites mirrored coding region sequence diversity across all species, as expected from the previous conservation analysis (Figure 3). Figure 4C shows a similar effect, with p53 cell-cycle–related sites displaying low sequence diversity. Intriguingly, the apoptotic-related p53 binding sites (Figure 4B) showed levels of sequence diversity that often met or exceeded those of the background promoter sequences. In rabbit and dog these apoptotic p53 binding sites have diverged so much that they may have lost function or could be under positive selection for mutation. The differences in positional sequence diversity between the two p53 RE subgroups were all highly significant (two-tailed paired t-test assuming unequal variances: dog = 1.4e–13, mouse = 4.3e–12, rabbit = 2.0e–17, and rat = 2.0e–15). These dramatic results indicate again that these different classes of p53 binding sites may be subject to widely dissimilar sequence constraints.

Discussion

The wide variety of genes transcriptionally regulated by p53 highlights the pleiotropic role of this master regulatory protein in many different biological pathways. Here, we addressed the conservation of human p53 binding sites across several mammals commonly used as experimental models. Examining global alignments of established human p53 binding sites, we found that common comparative genomics methods do not generally enhance p53 binding site prediction, although they can for NFκB, NRF2, and a subset of p53 target genes involved in cell-cycle regulation. This apparent lack of conservation for many functional human p53 binding sites suggests that regulation of the p53 response network may

Table 6. Efficacy of Mammalian Conservation to Predict Human TFBSs

TFBS	Area under ROC Curve[a]			
	Mouse	Rabbit	Rat	Dog
p53	0.696 (+/−0.041)	0.608 (+/−0.044)	0.674 (+/−0.042)	0.588 (+/−0.044)
p53, apoptosis	0.616 (+/−0.073)	0.469 (+/−0.075)	0.568 (+/−0.074)	0.437 (+/−0.074)
p53, cell cycle	0.783 (+/−0.060)	0.688 (+/−0.066)	0.798 (+/−0.058)	0.647 (+/−0.072)
NRF2	0.801 (+/−0.069)	0.666 (+/−0.084)	0.822 (+/−0.066)	0.672 (+/−0.083)
NFκB	0.852 (+/−0.060)	0.739 (+/−0.077)	0.802 (+/−0.069)	0.786 (+/−0.071)

[a] An AUC of 1 would represent a conservation metric that perfectly identified genuine TFBSs. Standard errors for the curve areas are shown.

doi:10.1371/journal.pgen.0030127.t006

be fine tuned for the needs of each species. By comparing sequence conservation between two p53 pathways, we have detected differences in the evolution of their regulatory elements. In particular, numerous functional human p53 REs in apoptosis-regulating sites, as well as the surrounding local sequence, show little homology to rodent sequences, suggesting that this ontology may have been shaped by primate-specific selection pressures that have resulted in turnover (loss or gain) of binding sites. This is supported to some degree by the very high mean sequence identity for all p53 REs between human and chimpanzee or monkey (Figure S1). However, turnover cannot be easily addressed by the global alignment method. For example, a short species-specific rearrangement such as an insertion of a repetitive DNA element (e.g., SINE, LINE, etc.) that contains a RE would not globally align and would show zero identity across species but might maintain functional response across species.

There were also seven DNA repair-regulating p53 REs in our dataset. Although there were not enough p53 RE sites to perform a statistically significant analysis, the average percent identity to the consensus sequence was similar to that of the apoptosis-related subset for human-to-rodent comparisons. These findings are significant considering the efforts to functionally model human p53 responses in the mouse (including cell cycle, apoptosis, and DNA repair) [33–35].

Complex molecular events (reviewed in [35]) regulate both p53 levels and activity prior to the transregulation of cell-cycle arrest and apoptosis genes. This results in large increases in p53 availability for binding to REs. Presumably the strength of p53 binding to a given target sequence has the effect of tuning regulation of the components of the p53 network within a species. Our data suggest that regulation of some p53 pathways, including apoptotic and DNA repair genes, may differ between humans and other mammalian species. Not only are REs in apoptotic and other genes different from cell-cycle genes in rodents, but they appear to differ from rabbits and dogs as well (Figure 4B). This unexpected excess of sequence diversity for apoptotic elements could be explained by recent positive selection in all of these species. Support for this comes from an evolutionary analysis of a functional, yet poor, consensus-matching p53 RE in the apoptotic gene *PIG3*. This study revealed that *PIG3* became p53 responsive only recently, during primate evolution [10] and is consequently only present and functional in apes and humans.

A recent emergence of primate-specific apoptosis p53 RE sites could explain the large number of interspecific differences identified following alignment to their orthologous mouse, rat, rabbit, or dog sites (Figure 4). Dermitzakis and Clark [36] observed a similar phenomena while surveying a broad number of TFBSs and concluded that a large percentage of apparently functional human sites were not functional in rodents (and vice versa). The authors suggest that loss and gain of TFBSs has been commonplace in both rodents and humans. On the other hand, the p53 protein itself has changed very little between species. Mouse and human p53 proteins are 85% identical and show equivalent transactivation of human apoptotic and cell-cycle REs in a yeast-based system [37]. The DNA binding domain of the p53 protein has near 100% homology across all mammalian species indicating strong purifying selection to maintain DNA binding function. Our data indicate that cell-cycle REs are also being maintained by purifying selection, while the evidence suggests that divergent positive selection has occurred among REs of apoptosis genes. The evolution of apoptosis-related p53 binding sites has strong biological plausibility, as it seems likely that such modifications could profoundly affect how a species responds to environmental stress and cellular damage. With exposure to DNA-damaging agents being a common environmental feature throughout mammalian history, selection pressure and the evolution of systems to maintain genome stability could be quite different in rodents and primates. For example, it was recently shown that the *Spalax* (mole rat), which lives its entire life underground, has a p53 protein with a very limited ability to induce several well-known human apoptotic genes in reporter assays. It is, however, quite capable of transactivating cell-cycle arrest genes [38]. The adaptation of the *Spalax* p53 response to a dramatically different environment underscores how separate pathways jumpstarted by the same transcription factor can have distinct evolutionary signatures.

Cross-species analysis of p53-regulated genes in relation to biological function is largely absent. Thus it is unknown whether any preservation of functionality in apoptosis-related p53 binding sites exists, or if divergence and positive selection have created uniquely primate response characteristics. We are currently evaluating how p53 RE variation across species affects binding and transactivation in a functional model system and have observed that some weak binding REs show high conservation (D. A. Bell, unpublished data). Other aspects of p53 pathways may evolve, such as the proteins that regulate the availability or quantity of active p53 protein, so that the sequence and binding affinity of affiliated binding sites could be coevolving with such changes.

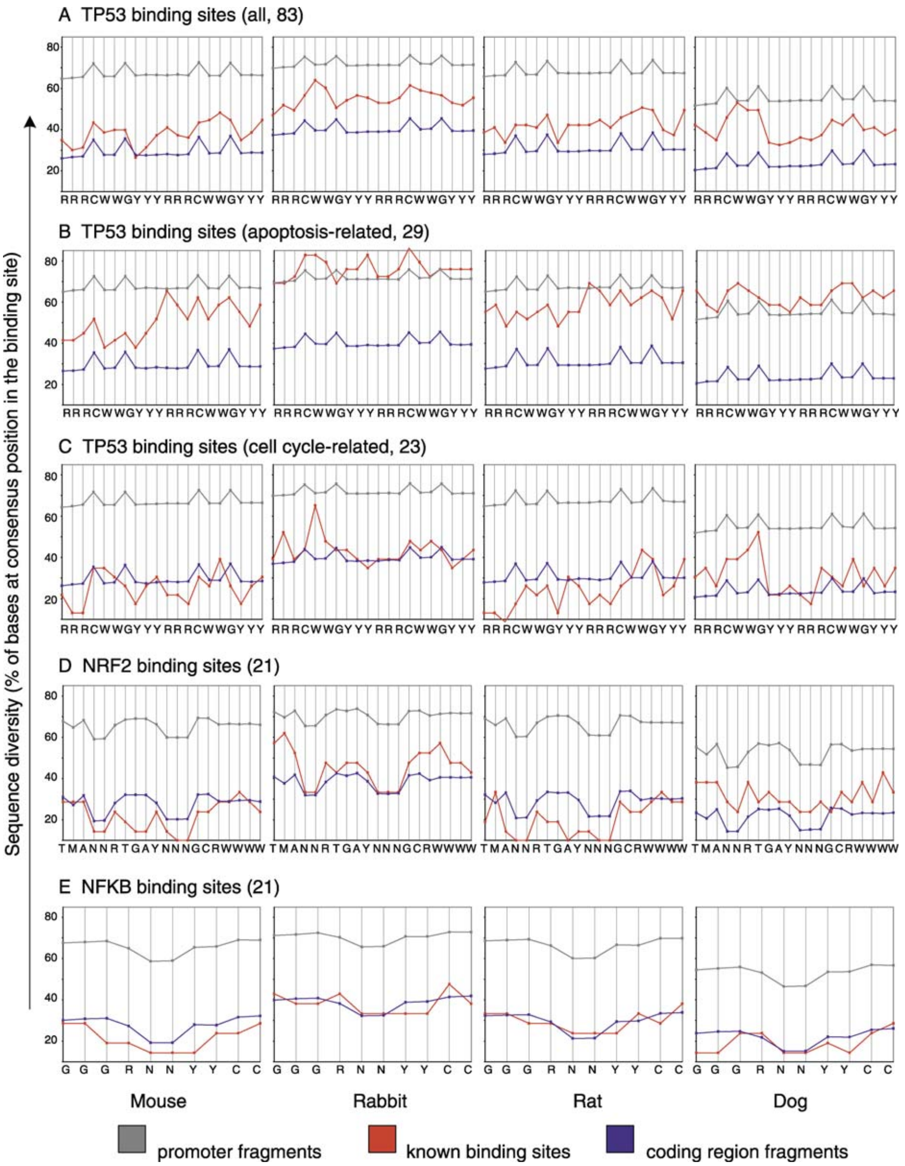


Figure 4. Positional Diversity of Human TFBSs Compared to Four Species

(A–E) Sequence diversity is calculated as the percentage of aligned bases at a certain position that vary from the accepted transcription factor consensus sequence, which is listed along the x-axis. The diversity levels of randomly chosen promoter sequences (gray) and randomly chosen coding region (blue) sequences (1,000 trials, error bars are narrower than the data points) are plotted so that level of genuine TFBS diversity (red, known sites) can be viewed in a genomic context. The shape of each plot reflects the level of degeneracy of the consensus at each nucleotide position (e.g., gray line, the p53 consensus has C and G at fixed positions creating peaks in the samples from the promoter)

doi:10.1371/journal.pgen.0030127.g004

This study makes several suggestions for computational analysis of p53 REs and regulatory sequence in general. Since these binding sites seem to be experiencing much short-term evolution and turnover, comparative genome analysis in a panel of old and new world monkeys, “phylogenetic shadowing,” may be a promising direction to enhance prediction accuracy [1]. Secondly, among those p53 REs exhibiting high conservation, mutations or polymorphisms that alter such sequences may be significant [35]. A key practical point is that if comparative genomics methods are used to identify putative functional regulatory regions, one should ensure that the choice of comparative species data is relevant to the selection pressure on the motifs of interest. For the p53 pathway, predictions based on mouse- or rat-to-human will not only generate a large excess of false positives, but many bona fide REs will be missed.

Overall, conservation analysis may be a convenient measuring stick for regulatory element function, but we have shown that it must be used with caution and may depend on the TFBS category being analyzed. A reduction in evolutionary conservation in p53 regulatory elements is likely due to species-specific selective pressures acting on the distinct biological differences among p53-regulatory pathways.

Materials and Methods

TFBSs with experimental support were located in the literature (see Table S1). Genome coordinates (National Center for Biotechnology [NCBI] 35.1, May 2004 release) were located using BLAT [39] searches against the human genome within the UCSC genome browser (Table S2) [31,40]. If a TFBS could not be found in the genome, it was removed from analysis, which left 83, 21, and 21 binding sites for p53, NRF2, and NFκB, respectively. The UCSC "multiz17way" conservation track [31,40,41] provided a 17-way multiple sequence alignment between current releases of the *Homo sapiens* genome and eleven additional mammals: *Pan troglodytes* (chimpanzee, November 2003), *Canis familiaris* (dog, May 2005), *Mus musculus* (mouse, May 2004), *Rattus norvegicus* (rat, June 2003), *Macaca mulatta* (rhesus monkey, January 2006), *Monodelphis domestica* (opossum, June 2005), *Bos taurus* (cow, March 2005), *Echinops telfairi* (tenrec, July 2005), *Loxodonta africana* (elephant, May 2005), *Oryctolagus cuniculus* (rabbit, May 2005), and *Dasypus novemcinctus* (armadillo, May 2005). This alignment set was used to find the corresponding locations of each TFBS within each genome. The accuracy of these alignment regions were manually inspected and verified by both confirming similar local gene organization as well as referencing independently generated paired human–mammal alignments (UCSC tables netMm7, netMonDom1, netBosTau2, netRn4, netCanFam2, netRheMac2, and netPanTro1) [42,43]. Similarly, a random list of promoter and cDNA sequences were obtained from Ensembl (http://www.ensembl.org) by referencing a genome coordinate list of all known human protein-encoding genes (version 35.1) with an Ensembl gene identifier [44]. For each gene, a coordinate range of length equal to the TFBS of interest was randomly picked in the (a) 3,500 bases 5′ to the gene start site and (b) within protein coding DNA sequence. Sites from these two lists of coordinates were randomly chosen to form a set of genome regions with the same number of members as each TFBS category: NRF2 (21), NFκB (21), p53 (83), p53 apoptosis (29), and p53 cell cycle/cell growth (23). This process was repeated 1,000 times with replacement to capture the variance in the data. As with the known TFBSs, the mammalian multiple alignment data from multiz17way were retrieved for each of these promoter and cDNA sites. The placement of target genes into p53 subcategories was based on the grouping used in Qian et al. and an annotated literature search (Table S1).

Pair-wise percent identities relative to the each TFBS consensus were calculated as the percentage of RE bases that were either (a) identical between the human and second genome or (b) mismatched but do not deviate from the consensus sequence (Figure 1). For p53 REs, the variable spacer region was not considered. The distribution of conservation for each TFBS set is shown in Figure 2. The conservation of the randomly chosen sets of coding region and promoter sequences, which represent the high and low extremes respectively of human genome conservation for TFBS comparisons, was calculated in the same fashion, and average results per x-axis bin for 1,000 trials are shown in Figure 2. *p*-Values describing the statistical difference between percent sequence identity means for each TFBS set (Table 4) were calculated using an unpaired two-tailed t-test with the assumption of unequal variances. The ORs in Table 4 were calculated as OR = ad/bc, where a = mean conservation of the column element, b = mean conservation of the row element, c = 100 − a, and d = 100 − b.

A list of CNC blocks, which represent the 5% most conserved portions of the human genome, were downloaded from the "most conserved" track of the UCSC genome browser. These were generated by constructing a phylogenetic two-state Hidden Markov Model [45] from the 17-way multiple alignment, which includes the human 35.1 genome release [31,41,44]. To generate the data in Table 5, the coordinates of each TFBS set were intersected with the CNC blocks. Likewise, the average number of randomly chosen promoter regions (1,000 trials) found within CNCs was determined. The probability of finding a greater fraction of CNCs in randomly chosen promoter regions over TFBS sites was calculated using a two-tailed binomial test.

For the ROC curves in Figure 3, the sensitivity (true positive rate) and 1-specificity (false positive rate) of TFBS prediction were calculated at each of 11 conservation thresholds. The true positive rate was the fraction of bona fide TFBSs with a consensus sequence percent identity above a given level. The false positive rate was calculated as the average number of sites (1,000 trials) that fell above a conservation threshold (consensus sequence percent identity) in an equivalently sized set of random promoter sequences. For example, 83 TFBSs composed the p53 ROC curve. Therefore, sets of 83 random promoter sites, where each site was 20 bases in length, were used to estimate the false positive rate. ROC AUCs (Table 6) and standard errors were calculated directly from the graphs using the trapezoid rule as described by Hanley and McNeil [46].

Positional sequence diversity for a TFBS, which is related to the inverse of sequence identity, was calculated as the percentage of human nucleotides (nt) at each position that had a nonconsensus mismatch or insertion/deletion event when compared to one of the four mammals (Figure 4). Each TFBS of a given class (i.e., TP53) and its alignment to a mammal (i.e., mouse) was pulled from the multiz multiple alignments to produce a set of paired sequence alignments. For each RE member of this pool, we counted the number of times where the first position of the mammalian sequence in the alignment differed from the human nt due to either (a) a nonconsensus mismatch or (b) an insertion/deletion event. This count was divided by the total number of TFBSs in a group (i.e., 83 for all p53 REs or 21 for all NFκB REs) to get the percentage of two-way alignments that differed at that position. This value is taken as the sequence diversity at that first position. This calculation was then performed for the remaining nts in the TFBS. Mismatches from human that did not alter the consensus motif did not increase the percent sequence diversity and were ignored.

Supporting Information

Figure S1. Distribution of Human-to-Primate TFBS Conservation Scores

Found at doi:10.1371/journal.pgen.0030127.sg001 (295 KB DOC).

Figure S2. ROC Curves for Four Human-to-Mammal Comparisons Indicate the Ability of Different Interspecific Comparisons to Identify Bona Fide TFBSs

Found at doi:10.1371/journal.pgen.0030127.sg002 (332 KB DOC).

Figure S3. Alignments for 52 p53 REs across Seven Mammalian Species Obtained from UCSC Multi17way Alignment Tool

Found at doi:10.1371/journal.pgen.0030127.sg003 (91 KB DOC).

Figure S4. Sequence Logos for Human p53 REs Used for Position Weight Matrix Model Construction

Found at doi:10.1371/journal.pgen.0030127.sg004 (152 KB DOC).

Table S1. Validated TFBSs Used in This Study and Supplementary References

Found at doi:10.1371/journal.pgen.0030127.st001 (454 KB DOC).

Table S2. Coordinates of Validated TFBSs

Found at doi:10.1371/journal.pgen.0030127.st002 (152 KB DOC).

Table S3. Correlation between RE Identity Distributions and Distributions Obtained from Sampling of Promoter Fragments

Found at doi:10.1371/journal.pgen.0030127.st003 (39 KB DOC).

Acknowledgments

The authors would like to thank Daniel Tomso, Maher Noureddine, Michelle Campbell, and Gary Pittman of the National Institute of Environmental Health Sciences (NIEHS) for helpful discussions. Krystal Finney (NIEHS) catalogued p53 RE information. Ben van Houten and Kellen Meadows of NIEHS and anonymous PloS reviewers provided helpful suggestions on the manuscript.

Author contributions. MMH, MAR, and DAB conceived and designed the experiments and wrote the paper. MMH performed the experiments. MMH, XW, and DAB analyzed the data. MMH and MAR contributed reagents/materials/analysis tools.

Funding. This research was entirely funded by the Intramural Research Program of the National Institutes of Health and NIEHS.

Competing interests. The authors have declared that no competing interests exist.

References

1. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, et al. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299: 1391–1394.
2. Jones NC, Pevzner PA (2006) Comparative genomics reveals unusually long motifs in mammalian genomes. Bioinformatics 22: e236–e242.
3. Dickmeis T, Muller F (2005) The identification and functional characterisation of conserved regulatory elements in developmental genes. Brief Funct Genomic Proteomic 3: 332–350.

4. Zhang Z, Gerstein M (2003) Of mice and men: Phylogenetic footprinting aids the discovery of regulatory elements. J Biol 2: 11.
5. Berg J, Willmann S, Lassig M (2004) Adaptive evolution of transcription factor binding sites. BMC Evol Biol 4: 42.
6. Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, et al. (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet 38: 223–227.
7. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, et al. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3: e170. doi:10.1371/journal.pbio.0030170
8. Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, et al. (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302: 1960–1963.
9. Wolfe KH, Sharp PM, Li WH (1989) Mutation rates differ among regions of the mammalian genome. Nature 337: 283–285.
10. Contente A, Zischler H, Einspanier A, Dobbelstein M (2003) A promoter that acquired p53 responsiveness during primate evolution. Cancer Res 63: 1756–1758.
11. Wang X, Tomso DJ, Liu X, Bell DA (2005) Single nucleotide polymorphism in transcriptional regulatory regions and expression of environmentally responsive genes. Toxicol Appl Pharmacol 207: 84–90.
12. Tomso DJ, Inga A, Menendez D, Pittman GS, Campbell MR, et al. (2005) Functionally distinct polymorphic sequences in the human genome that are targets for p53 transactivation. Proc Natl Acad Sci U S A 102: 6431–6436.
13. Hollstein M, Shomer B, Greenblatt M, Soussi T, Hovig E, et al. (1996) Somatic point mutations in the p53 gene of human tumors and cell lines: Updated compilation. Nucleic Acids Res 24: 141–146.
14. Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, et al. (2006) A global map of p53 transcription-factor binding sites in the human genome. Cell 124: 207–219.
15. el-Deiry WS, Kern SE, Pietenpol JA, Kinzler KW, Vogelstein B (1992) Definition of a consensus binding site for p53. Nat Genet 1: 45–49.
16. Harms K, Nozell S, Chen X (2004) The common and distinct target genes of the p53 family transcription factors. Cell Mol Life Sci 61: 822–842.
17. Wong J, Li PX, Klamut HJ (2002) A novel p53 transcriptional repressor element (p53TRE) and the asymmetrical contribution of two p53 binding sites modulate the response of the placental transforming growth factor-beta promoter to p53. J Biol Chem 277: 26699–26707.
18. Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M (2002) A polymorphic microsatellite that mediates induction of PIG3 by p53. Nat Genet 30: 315–320.
19. Qian H, Wang T, Naumovski L, Lopez CD, Brachmann RK (2002) Groups of p53 target genes involved in specific p53 downstream effects cluster into different classes of DNA binding sites. Oncogene 21: 7901–7911.
20. Dixit V, Mak TW (2002) NF-kappaB signaling. Many roads lead to madrid. Cell 111: 615–619.
21. Ghosh S, May MJ, Kopp EB (1998) NF-kappa B and Rel proteins: Evolutionarily conserved mediators of immune responses. Annu Rev Immunol 16: 225–260.
22. Pierce JW, Lenardo M, Baltimore D (1988) Oligonucleotide that binds nuclear factor NF-kappa B acts as a lymphoid-specific and inducible enhancer element. Proc Natl Acad Sci U S A 85: 1482–1486.
23. Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, et al. (2003) Distribution of NF-kappaB-binding sites across human chromosome 22. Proc Natl Acad Sci U S A 100: 12247–12252.
24. Wang XW, Tan NS, Ho B, Ding JL (2006) Evidence for the ancient origin of the NF-kappaB/IkappaB cascade: Its archaic role in pathogen infection and immunity. Proc Natl Acad Sci U S A 103: 4204–4209.
25. Wasserman WW, Fahl WE (1997) Functional antioxidant responsive elements. Proc Natl Acad Sci U S A 94: 5361–5366.
26. Wang X, Tomso DJ, Chorley BN, Cho HY, Cheung VG, et al. (2007) Identification of polymorphic antioxidant response elements in the human genome. Hum Mol Genet 16: 1188–1200.
27. Cho HY, Reddy SP, Kleeberger SR (2006) Nrf2 defends the lung from oxidative stress. Antioxid Redox Signal 8: 76–87.
28. Nioi P, McMahon M, Itoh K, Yamamoto M, Hayes JD (2003) Identification of a novel Nrf2-regulated antioxidant response element (ARE) in the mouse NAD(P)H:quinone oxidoreductase 1 gene: Reassessment of the ARE consensus sequence. Biochem J 374: 337–348.
29. Ramos-Gomez M, Kwak MK, Dolan PM, Itoh K, Yamamoto M, et al. (2001) Sensitivity to carcinogenesis is increased and chemoprotective efficacy of enzyme inducers is lost in nrf2 transcription factor-deficient mice. Proc Natl Acad Sci U S A 98: 3410–3415.
30. Kobayashi M, Itoh K, Suzuki T, Osanai H, Nishikawa K, et al. (2002) Identification of the interactive interface and phylogenic conservation of the Nrf2-Keap1 system. Genes Cells 7: 807–820.
31. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, et al. (2006) The UCSC genome browser database: Update 2006. Nucleic Acids Res 34: D590–D598.
32. Schneider TD, Stephens RM (1990) Sequence logos: A new way to display consensus sequences. Nucleic Acids Res 18: 6097–6100.
33. Hasty P, Campisi J, Hoeijmakers J, van Steeg H, Vijg J (2003) Aging and genome maintenance: Lessons from the mouse? Science 299: 1355–1359.
34. Hoeijmakers JH (2001) Genome maintenance mechanisms for preventing cancer. Nature 411: 366–374.
35. Toledo F, Wahl GM (2006) Regulating the p53 pathway: In vitro hypotheses, in vivo veritas. Nat Rev Cancer 6: 909–923.
36. Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover. Mol Biol Evol 19: 1114–1121.
37. Resnick MA, Inga A (2003) Functional mutants of the sequence-specific transcription factor p53 and implications for master genes of diversity. Proc Natl Acad Sci U S A 100: 9934–9939.
38. Ashur-Fabian O, Avivi A, Trakhtenbrot L, Adamsky K, Cohen M, et al. (2004) Evolution of p53 in hypoxia-stressed Spalax mimics human tumor mutation. Proc Natl Acad Sci U S A 101: 12236–12241.
39. Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12: 656–664.
40. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, et al. (2003) The UCSC genome browser database. Nucleic Acids Res 31: 51–54.
41. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14: 708–715.
42. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, et al. (2003) Human-mouse alignments with BLASTZ. Genome Res 13: 103–107.
43. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D (2003) Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A 100: 11484–11489.
44. Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, et al. (2005) Ensembl 2005. Nucleic Acids Res 33: D447–D453.
45. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034–1050.
46. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143: 29–36.