Cluster analysis of p53 binding site sequences reveals subsets with different functions

Cluster analysis of p53 binding site sequences reveals subsets with different functions

Ji-Hyun Lim[1,2,3], Natasha S. Latysheva[1,4], Richard D. Iggo[2,5] and Daniel Barker[1,6,*]

[1]School of Biology, University of St Andrews, St Andrews, Fife, KY16 9TH, UK.

[2]School of Medicine, University of St Andrews, Fife, KY16 9TF, UK.

[3]Current address: Alacris Theranostics GmbH, Fabeckstraße 60-62, D-14195 Berlin, Germany.
Email: j.lim@alacris.de

[4]Current address: MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0QH, UK.
Email: natashal@mrc-lmb.cam.ac.uk

[5]INSERM Unit U1218, University of Bordeaux, Institut Bergonie, 229 Cours de l'Argonne, 33076 Bordeaux, France.
Email: R.Iggo@bordeaux.unicancer.fr

[6]Current address: Institute of Evolutionary Biology, University of Edinburgh, Charlotte Auerbach Road, The Kings Buildings, Edinburgh, EH9 3FL, UK.
Email: Daniel.Barker@ed.ac.uk

*To whom correspondence should be addressed.
Email: Daniel.Barker@ed.ac.uk
Tel: + 44 (0)131 651 7812.

ABSTRACT

p53 is an important regulator of cell cycle arrest, senescence, apoptosis and metabolism, and is frequently mutated in tumours. It functions as a tetramer, where each component dimer binds to a decameric DNA region known as a response element. We identify p53 binding site subtypes and examine the functional and evolutionary properties of these subtypes. We start with over 1700 known binding sites and, with no prior labelling, identify two sets of response elements by unsupervised clustering. When combined they give rise to three types of p53 binding site. We find that probabilistic and alignment-based assessments of cross-species conservation show no strong evidence of differential conservation between types of binding site. In contrast, functional analysis of the genes most proximal to the binding sites provides strong bioinformatic evidence of functional differentiation between the three types of binding site. Our results are consistent with recent structural data identifying two conformations of the L1 loop in the DNA binding domain, suggesting that they reflect biologically meaningful groups imposed by the p53 protein structure.

KEYWORDS

p53, transcription factor, protein-DNA interaction, DNA sequence, cluster analysis, function, conservation, human genome

INTRODUCTION

The p53 transcription factor is well known for its role in suppressing tumour formation. The wild type form regulates transcription of genes implicated in cell cycle control, apoptosis and senescence.[1] Common oncogenic p53 mutants either induce a loss of these tumour suppressor functions or acquire properties that promote cell proliferation, invasion, and metastasis.[2,3] However, it is increasingly recognised that p53 has a plethora of functions mediated by a wide range of target genes, often with little or no connection to its 'classical' roles in cell cycle control and cell death.[4] These functions include metabolic reprogramming, stem cell maintenance, autophagy, and response to oxidative stress.[5,6] There are perhaps 300-3000 functional p53 binding sites in the human genome.[7,8,9] p53 binds to these sites as a homotetrameric 'dimer of dimers', where each dimer interacts with a redundant, approximately palindromic, decameric DNA motif called the p53 response element (RE).[10,11,12,13,14] The two

REs that bind to a full tetramer are either directly adjacent or separated by a few base pairs.[4,15,16]

The best characterised p53 REs are typically found either near the promoters or in the first introns of target genes[17] and are approximately summarized by the 10-base pattern RRRCWWGYYY[15], where 'R' indicates A or G, 'W' indicates A or T and 'Y' indicates C or T. In the ambiguous positions, not all residues are equally frequent; furthermore, other sequence variations exist. This flexibility suggests the hypothesis that different types of RE could mediate different biological processes, regulated by p53 with different binding specificities due to variable intrinsic sequence affinities,[18,19,20] different post-translational modifications, or by being in complex with different cofactors. Different biological functions might be expected to be subject to different strengths of natural selection, leading to varying rates of evolution of the associated REs. Indeed, it has been suggested that REs involved in apoptosis and DNA repair are more poorly conserved across species than those involved in the cell cycle.[21]

Here, we computationally investigate the existence of subsets of p53 binding site. One could divide p53 binding sites or REs into subsets based on criteria such as Gene Ontology (GO) annotation of the nearest gene,[22] and summarize the properties of these subsets. However, GO – though an important guideline in broad studies of function – reflects a human-imposed classification of function, is incomplete, and for intergenic binding sites may involve an arbitrary decision as to which of the two nearest genes is regulated by the site. Instead of beginning with GO-based subsets, we begin with the DNA sequences of known binding sites. In an unsupervised clustering procedure, we classify these on the basis of the sequence similarity of their constituent decameric REs. This allows groups of binding sites to emerge based on their sequence, without imposing any limitations based on possible functional consequences. Our procedure also removes the arbitrary effect of the strand of DNA considered. Once formed on the basis of sequence similarity, we investigate the function of binding site groups, using both GO annotation and cross-species conservation, on the assumption that groups differing in one or both of these respects may have functional significance.

We use this procedure to group the decameric REs into two clusters, arbitrarily labelled 'cluster 1' and 'cluster 2'. Then, given that two REs form a full p53 binding site, three groups of full binding site are possible: group '1,1' binding sites, consisting of two REs of cluster 1; group '2,2' binding sites, consisting of two REs of cluster 2; and group '1,2' binding sites, consisting of one RE of each type. We find evidence of functional differentiation between these binding site groups, but find no strong evidence of differential evolutionary conservation.

MATERIALS AND METHODS

Input data

We obtained 1757 p53 binding sites from the literature, as described by Lim et al.[23] These consist of 327 binding sites from Wei et al.,[1] 1422 from Smeenk et al.[7] after excluding a further 123 also present in Wei et al., and eight from Horvath et al.[21]. These 1757 binding sites are given in Supplementary Material.

Clustering p53 response elements

Within a binding site, we label as 'first' the response element (RE) that is nearer to the start of the chromosome in the conventional representation; it is thus an arbitrary property of the strand of the chromosomal sequence being considered. Binding sites were then each split into the two constituent REs, excluding any spacer. To ensure comparable bases were aligned, the 'second' RE was reverse-complemented. All REs were then represented as strings of bases from the base outermost in the binding site (5') on the left, to the innermost base (3') on the right. Redundant sequences were removed, leaving 1724 unique p53 RE sequences (Supplementary Material).

A symmetrical matrix of RE-to-RE Hamming distance was calculated[24]. Exploratory hierarchical clustering of this distance matrix with the unweighted paired-groups method using arithmetic averages (UPGMA)[25] produced varying results when repeated, presumably due to the arbitrary resolution of ties during the clustering procedure[26,27]. For the final clusters presented in this paper, we instead clustered using Ward's method,[28] which minimises an objective function at each stage in the clustering procedure. In typical implementations, the objective function is within-cluster variance, requiring Euclidian distances as input. Before clustering, we transformed the RE-to-RE Hamming distance matrix to Euclidian distance using the 'lingoes' function of the 'ade4' package[29] in R (http://www.r-project.org). Clustering with Ward's method was then performed using the 'hclust' function of R.

To divide the REs into subgroups, we drew a phenon line[30] on the Ward's method cluster diagram at a position that split the REs into two sets (i.e. k = 2 clustering). These two primary clusters of REs represent the most inclusive subsets supported by our analysis. We labelled these primary clusters of REs as 'cluster 1' and 'cluster 2'.

The robustness of the grouping of REs into primary clusters was assessed using a jackknife procedure. 1000 subsamples (jackknife replicates), each with a random set of 37% REs omitted,[31] were generated from the set of 1724 non-redundant p53 RE sequences. Hence, each replicate consists of a random subset of 1086 REs (63% of the set of nonredundant REs), sampled without replacement. Using the same procedure as for the analysis of the set of 1724 non-redundant REs, we clustered REs of each replicate at $k = 2$. We mapped each of the two clusters from each replicate to one of the primary clusters from the analysis of the full set of non-redundant REs. The replicate cluster with the highest proportion of overlap with cluster 1 of the primary clusters was mapped to primary cluster 1, and the other was mapped to primary cluster 2. As an indication of robustness of the clustering of the 1724 nonredundant REs, a G-test was used to investigate the correspondence between the assignment of REs to primary clusters in each jackknife replicate, and the assignment to the primary clusters in the analysis of the full, non-redundant set of 1724 REs.

To investigate the evolutionary relationships of the primary clusters of RE, position weight-matrices (PWMs) for the RE clusters were compared to known PWMs for p53, p63 and p73 REs from the Transfac database (BioBase Corporation; http://www.biobase-international.com/product/transcription-factor-binding-sites). If presented in Transfac as counts, binding site PWMs were converted to a frequency representation. Then, frequencies for each base position within the RE were taken as the mean of the frequencies for the first RE and for the reverse-complement of the second RE within the binding site. The resulting RE PWMs represent base frequencies starting from the outermost base of the binding site on the left (5') to the innermost base (3') on the right. PWMs were visualized as logos using Weblogo[32] with the nonredundant sequences as input in the case of cluster 1 and cluster 2, and a synthetic set of 5000 simulated sequences matching the composition of each base position in the RE PWM in the case of PWMs based on Transfac. Similarities among the innermost 9 bases of REs (the outermost base was excluded due to its absence in the p73 PWM, M04503), were quantified using profile-profile alignment scores calculated as the sum of dot-product scores for the individual base positions[33,34] without adjusting for background frequencies.

Functional and evolutionary analysis of p53 binding site subtypes

Based on the primary cluster membership of the two constituent REs in the un-jackknifed cluster analysis, we defined three groups of full p53 binding site. Each binding site may be a '1,1' binding site, consisting of two REs from cluster 1; a '2,2' binding site, consisting of two REs from cluster 2; or a '1,2' binding site, consisting of one RE from each cluster. In the latter case, we make no distinction between binding sites in which the RE from cluster 1 comes 'first'

and those in which it comes 'second', since this distinction is arbitrary, depending only on which strand of the double helix is being considered.

To investigate differential pairing between RE clusters within binding sites, we performed a G-test for evidence of association between 'cluster 1' and 'cluster 2' REs within the full, redundant set of 1757 p53 binding sites.

To test for functional differences between the three groups of binding sites ('1,1', '1,2' and '2,2'), nearest genes were assigned to binding sites as described by Lim et al.[23] Enrichment for GO biological process terms was performed with PANTHER[35] (http://www.pantherdb.org; version 11.0, released 2016-07-15). To test for overlap with hallmark gene sets, Ensembl Gene 85 IDs were converted to GRCh38.7 Entrez Gene IDs with Biomart then compared to the h.all.v5.1.entrez.gmt hallmark gene set in the Molecular Signatures Database[36] (MSigDB v5.1, January 2016 release; http://software.broadinstitute.org/gsea/msigdb/annotate.jsp).

Conservation levels for the three sets of binding site were first investigated using PhastCons scores,[37] which quantify negative selection by using a hidden Markov model-based method to estimate the probability that each nucleotide in a multiple alignment forms part of a conserved sequence element. PhastCons conservations scores take into account the conservation of neighbouring bases, which makes PhastCons scores a natural choice for detecting stretches of conserved sequence, such as p53 binding sites. We obtained PhastCons scores which represent levels of conservation (ranging 0 to 1, where higher values indicate higher conservation) across the following ten primate species: *Homo sapiens* (genome assembly hg19), *Pan troglodytes* (panTro2), *Gorilla gorilla* (gorGor1), *Pongo abelii* (ponAbe2), *Macaca mulatta* (rheMac2), *Papio hamadryas* (papHam1), *Callithrix jacchus* (calJac1), *Tarsius syrichta* (tarSyr1), *Microcebus murinus* (micMur1), and *Otolemur garnetti* (otoGar1). The PhastCons scores for every p53 binding site (as the average across all constituent base pairs within the site) were extracted using the UCSC table browser function (http://genome.ucsc.edu/cgi-bin/hgTables). For comparison, a background level of conservation was estimated from a pre-calculated, genome-wide phastCons score set downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons46way/primates). Random segments of the human genome, for which PhastCons scores were available, were sampled with replacement 10,000 times. Lengths of these segments were sampled from an empirical distribution, estimated from the lengths of the known p53 binding sites. Conservation scores for the various binding site groups ('1,1', '1,2' and '2,2') and the background were compared using Kruskal-Wallis (KW) tests, a non-parametric equivalent of ANOVA.

Secondly, as an additional approach to test binding site conservation, alignments of genomic regions containing p53 binding sites were extracted using the Ensembl Perl API.[38] Genomic coordinates of p53 binding sites in the three groups were first converted to hg19 coordinates, and the evolutionary conservation of the binding sites was assessed by calculating average percentage identities in three types of alignments. The alignments used were, firstly, the LastZ-net[39] pairwise alignment of *Homo sapiens* (GRCh37) versus *Pan troglodytes* (CHIMP2.1.4); secondly, the EPO[40,41] multiple alignment of six primates (*Homo sapiens, Gorilla gorilla, Pan troglodytes, Pongo abelii, Macaca mulatta, Callithrix jacchus*); and thirdly, the EPO alignment of 15 eutherian mammals (*Homo sapiens, Gorilla gorilla, Pan troglodytes, Pongo abelii, Macaca mulatta, Callithrix jacchus, Mus musculus, Rattus norvegicus, Oryctolagus cuniculus, Equus caballus, Felis catus, Canis familiaris, Sus scrofa, Bos taurus, Ovis aries*).

Methods are further discussed in the Supplementary Material.

RESULTS

Clusters of p53 response elements and binding sites

Ward's method clustering of nonredundant p53 REs based on Euclidian distance led to primary clusters of size 410 and 1314, which we designate 'cluster 1' and 'cluster 2', respectively (Figure 1; Figure 2).

The spread of results among jackknife replicates is summarized in Table 1. Table 1 shows very strong evidence of association between the original classification of REs into two clusters and the classification of REs into two clusters in jackknife replicates. In the majority of jackknife replicates, REs are assigned to the same primary cluster as in the analysis of the un-jackknifed set of 1724 nonredundant REs (Figure S1). Hence, the two primary clusters (Figure 1) are based on a pervasive difference that is present throughout the dataset.

For the full set of 1757 binding sites, 140 were in group '1,1' (consisting of two REs from cluster 1); 687 were in group '1,2' (consisting of one RE from each cluster); and 930 were in group '2,2' (consisting of two REs from cluster 2). Given the relative sizes of cluster 1 and cluster 2, these counts are not statistically significantly different from expectations under a null hypothesis of independent assignment of RE clusters to binding sites ($G = 0.689$, df = 2, $p = 0.709$).

Comparison of response element clusters with existing PWMs

When compared to PWMs for REs from known p53, p63 and p73 binding sites derived from Transfac, our cluster 1 REs are most similar to the p53 RE, then to the p73 RE. Both of our RE clusters are most similar to the Transfac p53 RE, then to the p73 RE, and least similar to the p63 RE (Table 2). Cluster 1 and the Transfac-based PWM for the p53 RE show a stronger CCC homopolymer in the three bases innermost in the binding site than do cluster 2, the p63 RE or the p73 RE (Figure 2).

Functional analysis of binding site groups

To identify potential differences in the function of genes near the different classes of binding site, we measured the overlap with genes defining 50 hallmark biological processes in the Molecular Signatures Database (MSigDB)[36]. The hallmark most strongly associated with all three of our binding site groups was 'genes involved in p53 pathways and networks', confirming the validity of the approach (Table S1). The results for the other hallmarks are shown visually in Figure 3, with numerical details in Table S1. The main functional difference found between binding site groups is that group '2,2' is associated with a much broader set of functions. Group '1,1' is mainly associated with signal transduction pathways, particularly pro-survival and oncogenic pathways. Group '1,2' had an intermediate phenotype, functionally broader than group '1,1' but not as broad as group '2,2'. GO enrichment analysis confirmed that group '2,2' is associated with a much broader set of functions than the other two groups (Tables S2-S4). Based on these analyses, we conclude that a switch between '1,1' and '2,2' modes of DNA binding would change the spectrum of biological functions activated by p53.

Conservation of binding site groups

The conservation of binding sites in each group was first assessed using PhastCons scores, which are base-by-base probabilities of a given nucleotide belonging to an evolutionarily conserved element. The distributions of PhastCons scores for the three classes of binding sites, as well as the conservation scores across the length-matched genomic background, are shown in Figure 4. There is no statistically significant difference between conservation scores across the three groups of binding site (KW X^2 = 2.49, df = 2, p = 0.288). Conservation of binding sites and flanking regions was also assessed (Figure S2). No statistically significant differences in evolutionary conservation were found when sequences flanking the binding sites were included by adding 50 base pairs on each side of a binding site (forming ~110 bp regions, i.e. 100 bp flanking regions; KW X^2 = 0.052, df = 2, p = 0.974). Similarly, no statistically significant difference was found when longer, 1000 bp flanking regions were included (forming

~1010 bp regions; KW X^2 = 1.78, df = 2, p = 0.410). The difference between conservation scores for all p53 binding sites (mean = 0.176, median = 0.044) and background levels of genome conservation (mean = 0.127, median = 0.041) was also not statistically significant (KW X^2 = 0.100, df=2, p=0.752). Similarly, no statistically significant differences were found when separately comparing the conservation of each binding site to the background level of conservation.

The distribution of PhastCons conservation scores in both the p53 binding site and genomic background sequences appears slightly bimodal (Figure 4). The second peak, representing the highest observed conservation levels, is more pronounced for binding sites than for the genomic background. We find that 102 binding sites have PhastCons conservation scores greater than or equal to 0.90, representing 5.9% of all binding sites, but only 195 (2.0%) of length-matched background genomic regions fall into this highly conserved category. This constitutes strong evidence that binding sites may have a larger subset of highly conserved sequences (G-test vs genomic background as an extrinsic null hypothesis; G = 73.45, df = 1, $p < 2.2\text{x}10^{-16}$). Further examining the highly conserved p53 subset, we find that group '1,1' sites may be slightly overrepresented. Group '1,1' represents 8.1% of all binding sites, but constitutes 9.8% of the highly conserved subset, though this difference is not statistically significant (G-test on 2x2 contingency table; G = 0.42, df = 1, p = 0.52). Applying a less stringent (but high) conservation score cut-off of 0.8, 141 binding sites (8.2%) are above the cut-off, compared to the genomic background level of 2.9% (G-test vs extrinsic null hypothesis; G = 92.34, df = 1, $p < 2.2\text{x}10^{-16}$), and the proportion of the conserved subset included in group '1,1' rises to 12.1%, though this difference remains statistically non-significant (G-test on 2x2 contingency table; G = 2.93, df = 1, p = 0.087).

The finding of no strong evidence that p53 binding sites are more conserved than background genomic sequences is in accord with the observation that transcription factor binding sites show high evolutionary turnover, both in general[42] and particularly for p53.[21] There was no strong evidence of a difference in conservation between the functionally broader group '2,2' and the others (group '1,1' with group '1,2' mean = 0.177, median = 0.046; group '2,2' mean = 0.175, median = 0.041; KW X^2 = 0.429, df = 1, p = 0.512).

As an alternative means to analyse binding site conservation, three sets of multiple alignments were examined to study p53 binding site sequence divergence over increasingly long spans of evolutionary time (chimp-human, primate, and eutherian mammal; Figure S3). Overwhelmingly, these alignments support the PhastCons-based conclusion of no differential conservation between binding site groups (Table 3). The sole conservation differences close

to the conventional cut-off for statistical significance for a single test ($p < 0.05$) occur in the chimp-human comparison: group ‘1,1’ binding sites are more highly conserved between humans and chimps than both group ‘1,2’ ($p = 0.051$) and group ‘2,2’ ($p = 0.040$; Table 3). This may be taken as weak evidence for the conservation of group ‘1,1’ p53 binding site functionality between chimps and humans, or equivalently, the relative divergence of p53 binding sites related to non-canonical functions (i.e. those containing cluster 2 REs). However, the statistical significance is borderline and may be misleading due to multiple testing. Higher conservation of group ‘1,1’ binding sites was not observed in the primate alignments or in the mammal alignments (Table 3).

DISCUSSION

We have shown that subtle differences in p53 binding site functionality can be identified by clustering the constituent decamers on the basis of sequence similarity. We obtained a robust grouping of decamers into two major clusters. These two clusters of decamers can give rise to three groups of binding site, each composed of one of the three possible combinations of decamer. The frequencies of specific pairings of decamers from the two clusters into binding sites show no strong difference compared to random expectation, and we find no appreciable difference in conservation compared to background genome conservation levels. Furthermore, the three binding site groups also showed little evidence of differential conservation between themselves, with the strongest evidence hinting at relatively strong chimp/human conservation of group ‘1,1’ binding sites, though with only borderline statistical significance. However, we find that genes near ‘2,2’ sites have a much broader range of functions than genes near ‘1,1’ and ‘1,2’ sites, (Figure 3 and Tables S1-S4). Combined with the robustness of the RE clusters demonstrated by jacknifing, and with results from earlier studies (discussed below), we conclude that switching p53 from a '1,1' to a '2,2' mode of binding would substantially change the functional consequences of p53 activation.

Our results confirm a long-standing suspicion that p53 binding sites are not simply duplicated copies of a symmetrical RRRCWWGYYY decamer. Instead, the REs in cluster 1 are C-rich in the final three positions, which correspond to the innermost positions in the middle of a full 20mer (or larger) binding site. Because of the way we report the decamer sequences, ‘1,1’ binding sites will tend to have the sequence ‘CCCGGG’ at the centre of the 20mer. This is the sequence that was found in the original SELEX study that first defined the p53 binding site.[43] Shortly thereafter we showed that mutations in the L1 loop alter the affinity and specificity of DNA binding,[18] but an understanding of the mechanism had to wait until the Halazonetis group

discovered that the L1 loop in *C. elegans* p53 contains a small alpha helix.[11] They went on to show that the L1 loop in human p53 can form the same alpha helix.[44] The lysine 120 DNA contact residue lies at the tip of the loop. Accordingly, formation of the alpha helix retracts the lysine from the DNA. The discovery that the L1 loop can adopt two different conformations immediately suggests an explanation for the asymmetry in the cluster 1 and cluster 2 sequences in our study. The L1 loop is in the retracted form in the outer p53 subunits in the tetramer.[44,45,46] In this form lysine 120 can not reach into the major groove to contact the bases, so the sequence is less constrained. In contrast, the loop is in the extended form in the inner two subunits, allowing lysine 120 to forms hydrogen bonds with the bases in the major groove. The hydrogen bonds between the side chains of lysine 120, cysteine 277 and arginine 280 and the DNA are shown as yellow dotted lines in Figure 5A. The L1 loop is shown in the extended form in Figure 5B, and in the retracted state in Figure 5C. Switching to the extended conformation allows induced fitting of the protein to the DNA when the correct sequence is present.[44,45] It is likely that the L1 loop adopts many different conformations while searching for the correct sequence and that, thanks to induced fitting, this leads to important differences in the kinetics of binding that depend on the sequence.[44,45] In addition to the inner-outer asymmetry caused by changes in the conformation of the L1 loop there are differences between the hydrogen bonds formed, depending on the exact sequence at positions 8 and 9 in the decamer: cysteine 277 forms a hydrogen bond with either O4 of thymine or N4 of cytosine at position 8; lysine 120 forms hydrogen bonds with N7 and O6 of guanine but only N7 of adenine at position 9; and hydrophobic and van de Waals forces from alanine 276 and cysteine 277 stabilise the C5 methyl group in T at position 8.[10,44,45] Taken together, these data would lead us to expect p53 to bind with decreasing affinity to ‘1,1’, '1,2', and '2,2' sites. Hallmark analysis reveals a preference for prosurvival and oncogenic signalling pathways for ‘1,1’ sites (Figure 3). This is consistent with old suggestions that p53 promotes survival early after activation, and only binds to all of its targets if the signal persists and p53 accumulates. Originally this was interpreted as a binary switch between cell cycle arrest and apoptotic sites, with the latter containing only a single decamer[18,20] and having a lower affinity for p53 [20,19], but the multiplication of p53 functions over time means the effects are likely to be more diverse and to depend heavily on the cellular context. The most important DNA binding residue in p53 is arginine 280, which forms hydrogen bonds with the G base paired to the invariant C at position 4 in the pentamer. The corresponding positions in the decamer are 4 (C) and 7 (G). The pattern in cluster 1, with a stronger preference for G at position 7 than for C at position 4, is reminiscent of a binding site profile identified by Veprintsev and Fersht.[8] Interestingly, acetylation of lysine 120 [47,48] negated the difference.[49] In addition to acetylation of K120, the cell can manipulate the sequence specificity of p53 through multiple mechanisms, for example binding to Hzf and ASPP proteins.[50,51] Indeed, many publications have described plausible

regulatory mechanisms based on post-translational modifications and protein-protein interactions (reviewed by Carvajal and Manfredi[52]) that could explain the differences we have found by clustering of p53 binding sites. Given the elegant structural studies from the Halazonetis group cited above, we suspect that these regulatory mechanisms converge on the L1 loop and switch p53 from a '1,1' to a '2,2' mode of binding.

In conclusion, we have shown that p53 binding sites can be classified into groups that may reflect the different modes of DNA binding that have been described in structural studies. Integration of sequence-based clustering with data on post-translational modification, cofactor binding and changes in the structure of the DNA binding domain is a promising direction for future research.

ACKNOWLEDGEMENTS

We acknowledge the financial support of the University of St Andrews School of Medicine and a BBSRC Doctoral Training Grant [BB/D526845/1] (studentship to J.-H.L.); The University of St Andrews Undergraduate Research Internship Programme (award to N.L.); and the French National Research Agency [ANR grant ANR-08-CEXC-016-01 to R.I.].

AUTHOR CONTRIBUTIONS

Conceived and designed the analyses: J-HL, NSL, RDI, DB. Performed the analysis: J-HL, NSL, RDI, DB. Wrote the manuscript: J-HL, NSL, RDI, DB. Agree with manuscript results and conclusions: J-HL, NSL, RDI, DB. All authors read and approved of the final manuscript.

REFERENCES

1. Wei CL, Wu Q, Vega VB, et al. A global map of p53 transcription-factor binding sites in the human genome. *Cell*. 2006;124:207-19.

2. Muller PAJ, Vousden KH. p53 mutations in cancer. *Nat. Cell Biol*. 2013;15:2-8.

3. Muller PAJ, Vousden KH. Mutant p53 in cancer: new functions and therapeutic opportunities. *Cancer Cell*. 2014;25:304-317.

4. Menendez D., Inga A, Resnick MA. The expanding universe of p53 targets. *Nat. Rev. Cancer*. 2009;9:724-37.

5. Bieging KT, Mello SS, Attardi LD. Unravelling mechanisms of p53-mediated tumour suppression. *Nat. Rev. Cancer*. 2014;14:359-70.

6. Hager KM, Gu W. Understanding the non-canonical pathways involved in p53-mediated tumor suppression. *Carcinogenesis*. 2014;35:740-6.

7. Smeenk L, van Heeringen SJ, Koeppel, M., et al. Characterization of genome-wide p53-binding sites upon stress response. *Nucleic Acids Res*. 2008;36:3639-54.

8. Veprintsev DB, Fersht AR. Algorithm for prediction of tumour suppressor p53 affinity for binding sites in DNA. *Nucleic Acids Res*. 2008;36:1589-98.

9. Wang B, Niu D, Lam TH, Xiao Z, Ren EC. Mapping the p53 transcriptome universe using p53 natural polymorphs. *Cell Death Differ*. 2014;21:521-32.

10. Cho Y, Gorina S, Jeffrey PD, Pavletich NP. Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science*. 1994;265:346-55.

11. Huyen Y, Jeffrey PD, Derry WB, et al. Structural differences in the DNA binding domains of human p53 and its *C. elegans* ortholog Cep-1. *Structure*. 2004;12:1237-43.

12. Kitayner M, Rozenberg H, Kessler N, et al. Structural basis of DNA recognition by p53 tetramers. *Mol. Cell*. 2006;22:741-53.

13. Ho WC, Fitzgerald MX, Marmorstein R. Structure of the p53 core domain dimer bound to DNA. *J. Biol. Chem*. 2006;281:20494-502.

14. Joerger AC, Fersht AR. Structural biology of the tumor suppressor p53. *Annu. Rev. Biochem*. 2008;77:557-82.

15. El-Deiry WS, Kern SE, Pietenpol JA, Kinzler KW, Vogelstein B. Definition of a consensus binding site for p53. *Nat. Genet.* 1992;1:45-9.

16. Riley T, Sontag E, Chen P, Levine A. Transcriptional control of human p53-regulated genes. *Nat. Rev. Mol. Cell Biol*. 2008;9,402-12.

17. Brady CA, Attardi LD. p53 at a glance. *J. Cell Sci.* 2010;123:2527-2532.

18. Freeman J, Schmidt S, Scharer E, Iggo R. Mutation of conserved domain II alters the sequence specificity of DNA binding by the p53 protein. *EMBO J*. 1994;13:5393-400.

19. Ludwig RL, Bates S, Vousden KH. Differential activation of target cellular promoters by p53 mutants with impaired apoptotic function. *Mol. Cell. Biol.* 1996;16:4952-60.

20. Saller E, Tom E, Brunori M, et al. Increased apoptosis induction by 121F mutant p53. *EMBO J*. 1999;18:4424-4437.

21. Horvath MM, Wang X, Resnick MA, Bell DA. Divergent evolution of human p53 binding sites: cell cycle versus apoptosis. *PLoS Genet.* 2007;3:e127.

22. Gene Ontology Consortium. Gene ontology annotations and resources. *Nucleic Acids Res.* 2013;41:D530-5.

23. Lim J-H, Iggo RD, Barker D. Models incorporating chromatin modification data identify functionally important p53 binding sites. *Nucleic Acids Res*. 2013;41:5582-93.

24. Hamming RW. Error detecting and error correcting codes. *Bell System Techincal Journal*. 1950;29:147-60.

25. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*. 1958;28:1409-38.

26. Backeljau T, De Bruyn L, De Wolf H, Jordaens K, Van Dongen S, Winnepennickx W. Multiple UPGMA and neighbour-joining trees and the performance of some computer pacjkages. *Mol. Biol. Evol*. 1996;13:309-13.

27. Scherma D, Podani J, Erős T. Mesuring the contribution of community members to functional diversity. *Oikos*. 2009;118:961-71.

514 28. Ward JH. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*
515 1963;58:236-44.
516
517 29. Dray S, Dufour AB. The ade4 package: implementing the duality diagram for ecologists. *J.*
518 *Stat. Softw.* 2007;22:1-20.
519
520 30. Sneath PHA, Sokal RR. *Numerical taxonomy: The Principles and Practice of Numerical*
521 *Classification*. San Francisco: W.H. Freeman; 1973.
522
523 31. Farris JS, Albert VA, Källersjö M, Lipscomb D, Kluge AG. Parsimony jackknifing
524 outperforms neighbor joining. *Cladistics*. 1996;12:99-124.
525
526 32. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator.
527 *Genome Res*. 2004;14:1188-90.
528
529 33. Ohlson T, Wallner B, Elofsson A. Profile-profile methods provide improved fold-recognition:
530 a study of different profile-profile alignment methods. *Proteins*. 2004;57:188-97.
531
532 34. Wang G, Dunbrack RL. Scoring profile-to-profile sequence alignments. *Protein Science*.
533 2004;13:1612-26.
534
535 35. Mi H, Poudel S, Muruganujan A, Casagrande JT, et al. PANTHER version 10: expanded
536 protein families and functions, and analysis tools. *Nucleic Acids Res*. 2016;44(D1):D336-342.
537
538 36. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette M, Paulovich A,
539 Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-
540 based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA.
541 2005;102:15545-15550.
542
543 37. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate,
544 insect, worm, and yeast genomes. *Genome Res*. 2005;15:1034-50.
545
546 38. Stabenau A, McVicker G, Melsopp C, Proctor G, Clamp M, Birney E. The Ensembl core
547 software libraries. *Genome Res*. 2004;14:929-33.
548
549 39. Harris RS. Improved pairwise alignment of genomic DNA. Ph.D. Thesis, Pennsylvania
550 State University; 2007.

40. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res*. 2008;18:1814-28.

41. Paten B, Herrero J, Fitzgerald, S, et al. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res*. 2008;18:1829-43.

42. Doniger SW, Fay JC. Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput. Biol*. 2007;3:e99.

43. Funk WD, Pak DT, Karas RH, Wright WE, Shay JW. A transcriptionally active DNA-binding site for human p53 protein complexes. *Mol. Cell. Biol*. 1992;12:2866-71.

44. Petty TJ, Emamzadah S, Costantino L, et al. An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity. *EMBO J*. 2011;30:2167-76.

45. Emamzadah S, Tropia L, Halazonetis TD. Crystal structure of a multidomain human p53 tetramer bound to the natural CDKN1A (p21) p53-response element. *Mol. Cancer Res.* 2011;9:1493-9.

46. Emamzadah S, Tropia L, Vincenti I, Falquet B, Halazonetis TD. Reversal of the DNA-binding-induced loop L1 conformational switch in an engineered human p53 protein. *J. Mol. Biol.* 2014;426:936-44.

47. Sykes SM, Mellert HS, Holbert MA, Li K, Marmorstein R, Lane WS, McMahon SB. Acetylation of the p53 DNA-binding domain regulates apoptosis induction. *Mol. Cell*. 2006;24:841-51.

48. Tang Y, Luo J, Zhang W, Gu W. Tip60-dependent acetylation of p53 modulates the decision between cell-cycle arrest and apoptosis. *Mol. Cell*. 2006;24:827-39.

49. Arbely E, Natan E, Brandt T, et al. Acetylation of lysine 120 of p53 endows DNA-binding specificity at effective physiological salt concentration. *Proc. Natl. Acad. Sci. U.S.A.* 2011;108:8251-6.

50. Das S, Raj L, Zhao B, Kimura Y, Bernstein A, Aaronson SA, Lee SW. Hzf Determines cell

588 survival upon genotoxic stress by modulating p53 transactivation. *Cell.* 2007;130:624-37.
589
590 51. Samuels-Lev Y, O'Connor DJ, Bergamaschi D, Trigiante G, Hsieh JK, Zhong S, Campargue I,
591 Naumovski L, Crook T, Lu X. ASPP proteins specifically stimulate the apoptotic function of p53. *Mol.*
592 *Cell*. 2001;8:781-94.
593
594 52. Carvajal LA, Manfredi JJ. Another fork in the road—life or death decisions by the tumour
595 suppressor p53. *EMBO Rep*. 2013;14:414-21.
596

TABLES

Table 1. Contingency table showing the relationship between response element (RE) classification in the original cluster analysis (Figure 1) and re-classification in jackknife replicates. A highly statistically significant association was observed between the original classification of REs into two clusters and the classification of REs into two clusters in jackknife replicates (G = 519.98, d.f. = 1, $p < 2.2 \times 10^{-16}$).

Counts	**Replicate cluster 1**	**Replicate cluster 2**	**Totals**
Original cluster 1	440	70	510
Original cluster 2	344	870	1214
Totals	784	940	n = 1724

Table 2. Dot-product alignment scores between PWMs for RE cluster 1, RE cluster 2, and PWMs for the p53 RE, p63 RE and p73 RE derived from the Transfac database (M01651, M07138 and M04503). To match the PWM for p73, which has REs of length 9, the first (outermost) base of the other PWMs was omitted. The alignment score depends both on the extent of matching between profiles and the extent of ambiguity within profiles, and is not a metric. Scores are symmetrical and are only given for the bottom-left portion of the table. Scores can range from a maximum of 9, for two unambiguous 9-base PWMs which perfectly match, to a minimum of 0.

	Cluster 1	**Cluster 2**	**p53 Transfac**	**p63 Transfac**	**p73 Transfac**
Cluster 1	4.8	.	.	.	.
Cluster 2	4.3	4.7	.	.	.
p53 Transfac	4.9	5	5.5	.	.
p63 Transfac	4.3	4.5	4.8	4.5	.
p73 Transfac	4.5	4.9	5.3	4.7	5.3

Table 3. p53 binding site conservation as judged by averaged percentage identities from multiple sequence alignments. In each alignment, the mean and median percentage identities for the three binding site groups are shown. The distribution of percentage identities in each binding site group was pairwise tested against the remaining 2 groups (X^2 = Kruskal-Wallis X^2 statistic; p = p value).

	Binding site	**Mean**	**Median**	**Group '1,1'**	**Group '1,2'**
Chimp-human divergence	Group '1,1'	99.18	100	.	.
	Group '1,2'	98.70	100	X^2 = 3.82, p = 0.051	.
	Group '2,2'	98.60	100	X^2 = 4.21, p = 0.040	X^2 = 0.03, p = 0.866
Primate divergence	Group '1,1'	93.73	95	.	.
	Group '1,2'	92.07	95	X^2 = 0.85, p = 0.358	.
	Group '2,2'	92.46	95	X^2 = 0.97, p = 0.325	X^2 = 0.007, p = 0.935
Eutherian mammal divergence	Group '1,1'	82.52	82.37	.	.
	Group '1,2'	82.72	83.30	X^2 = 0.21 p = 0.645	.
	Group '2,2'	82.41	82.41	X^2 = 0.04, p = 0.842	X^2 = 1.56, p = 0.221

FIGURE CAPTIONS

Figure 1. Summary of dendogram obtained by cluster analysis of the 1724 nonredundant decamers. For visualization purposes, an arbitrary phenon line was drawn at a height of 38. The number of sequences in each resulting sub-cluster is shown, along with the logo summarizing those sequences, with bases ranging from 1 (outermost) to 10 (innermost) in the binding site. The logo *y*-axis represents information content, with ticks at 1 and 2 bits. The full dendogram is available as a file in Newick format in the Supplementary Material.

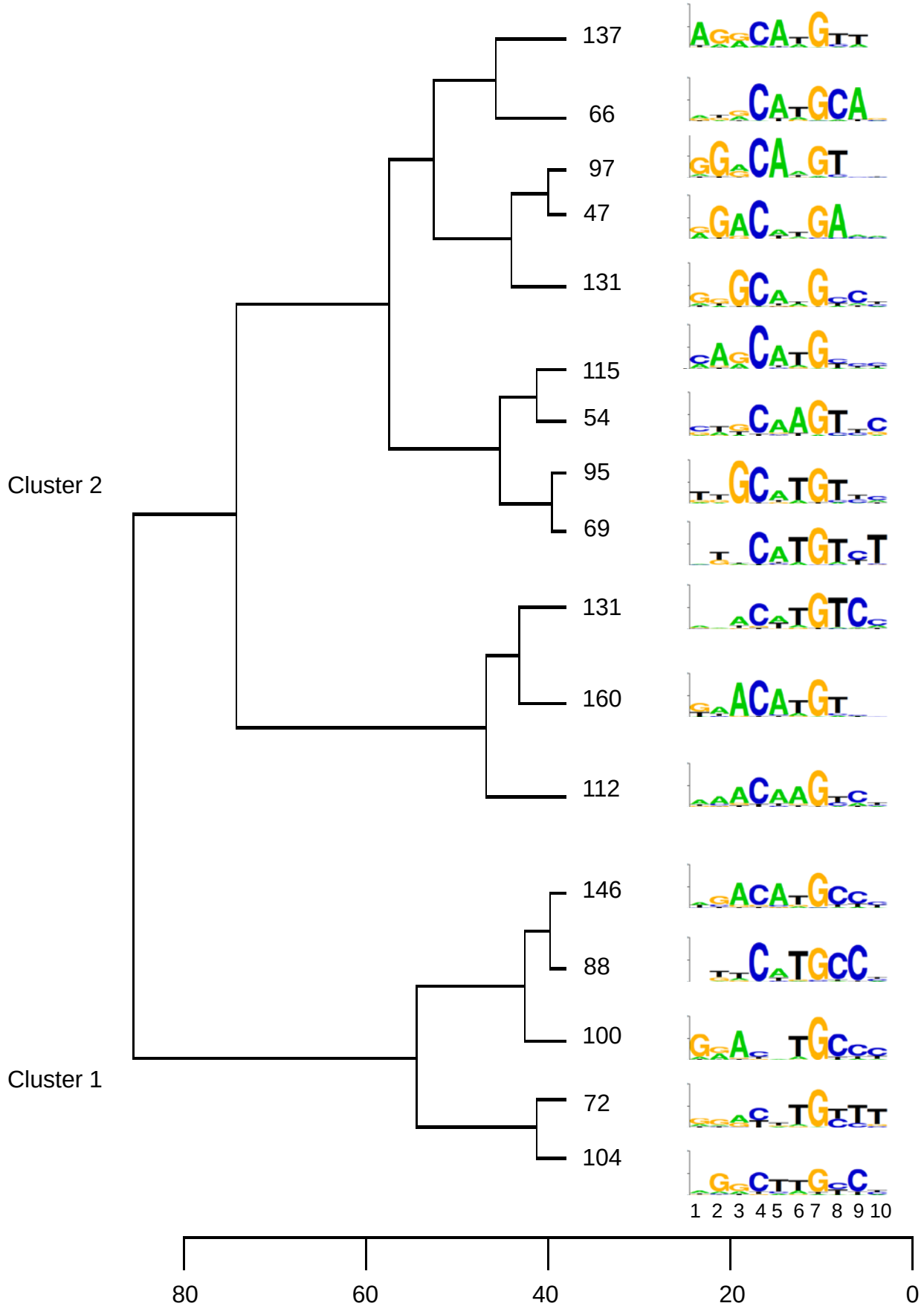
Figure 2. Sequence logos for (a) cluster 1 REs, (b) cluster 2 REs, (c) p53 Transfac RE, (d) p63 Transfac RE and (e) p73 Transfac RE. Bases range from 1 (outermost) to 9 or 10 (innermost) in the binding site. (c), (d) and (e) are based on Transfac M01651, M07138 and M04503, respectively.

Figure 3. Functional enrichment for hallmark biological processes. The genes nearest to the binding sites were used to create putative target gene lists for each group. The bars in the figure show the relative enrichment for genes in each hallmark; the dots show the *p*-value expressed as -log10. Only hallmarks for which at least one group gave $p < 0.0001$ are shown; within each hallmark missing bars correspond to associations with $p > 0.01$. For numerical details see Table S1. The terms in MSigDB corresponding to the labels are: PI3K, PI3K_AKT_mTOR_signaling; NFKB, TNFA_signaling_via_NFKB; Hyp, hypoxia; TOR, mTORC1_signaling; RAS, KRAS_signaling_up; UV, UV_response_down; Apo, apoptosis; EMT, epithelial_mesenchymal_transition; ER, estrogen_response_early; Inf, inflammatory_response; Myo, myogenesis; Gly, glycolysis; IL2, IL2_STAT5_signaling; Xeno, xenobiotic_metabolism.

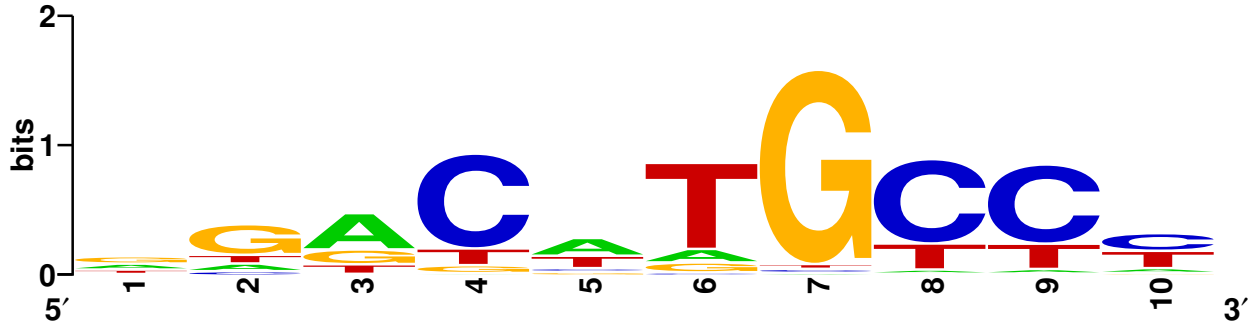
Figure 4. Histograms of PhastCons evolutionary conservation scores for binding sites in our p53 binding site group ‘1,1’ (n = 140), group ‘1,2’ (n = 687), group ‘2,2’ (n = 930) and the genomic background (n = 10,000), across 10 species of primates. Dashed lines indicate means for each group.

Figure 5. p53 DNA binding. (A) The p53 loop-sheet-helix is shown in contact with the major groove of the DNA. Amino acid 120K (cyan) binds to G on the Watson strand; 277C (orange) binds to T and 280R (red) to G on the Crick strand. Amino acid 120K arises from the tip of the L1 loop (the green line at the bottom of the figure). Hydrogen bonds are shown as dotted yellow lines. (B) The L1 loop in the extended form, as in panel (A). (C) The L1 loop in the retracted

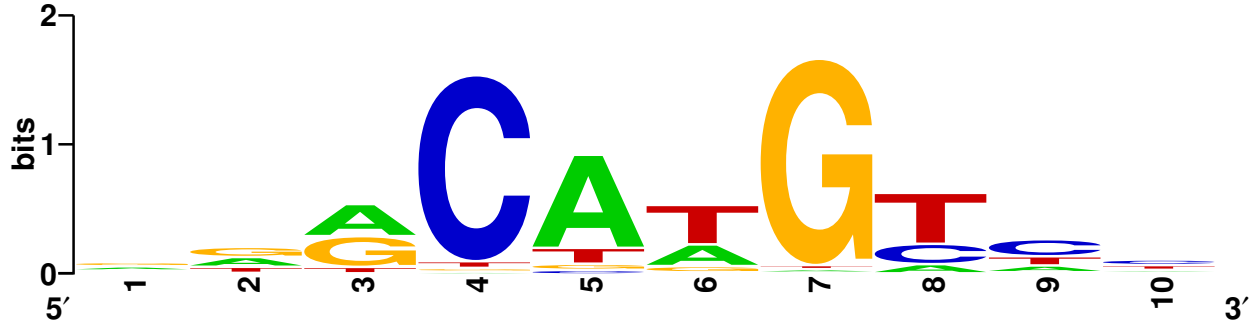
663 form. The figures were made with PyMOL (Schrödinger, LLC) from PDB structure 3Q05; for a
664 detailed description of the p53 DNA-protein interaction see [44,45,46].


137
66
97
47
131
115
54
95
69
131
160
112
146
88
100
72
104
Cluster 2
Cluster 1
1 2 3 4 5 6 7 8 9 10
80
60
40
20
0

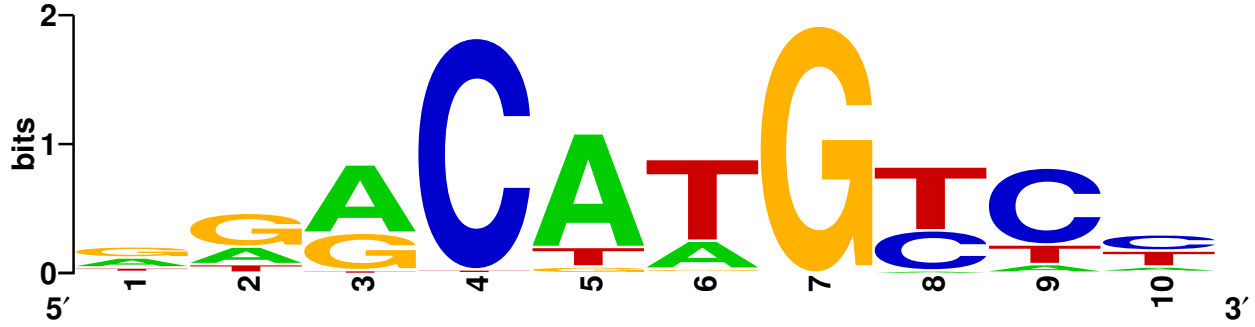
(A) Cluster 1



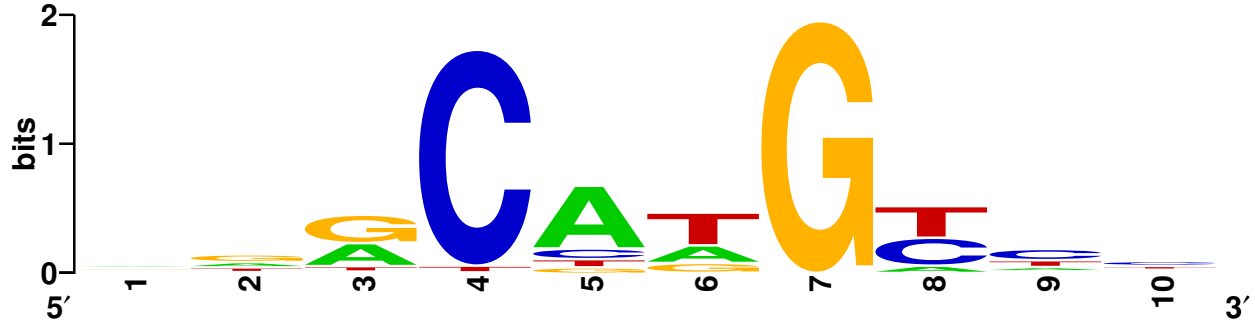
(B) Cluster 2



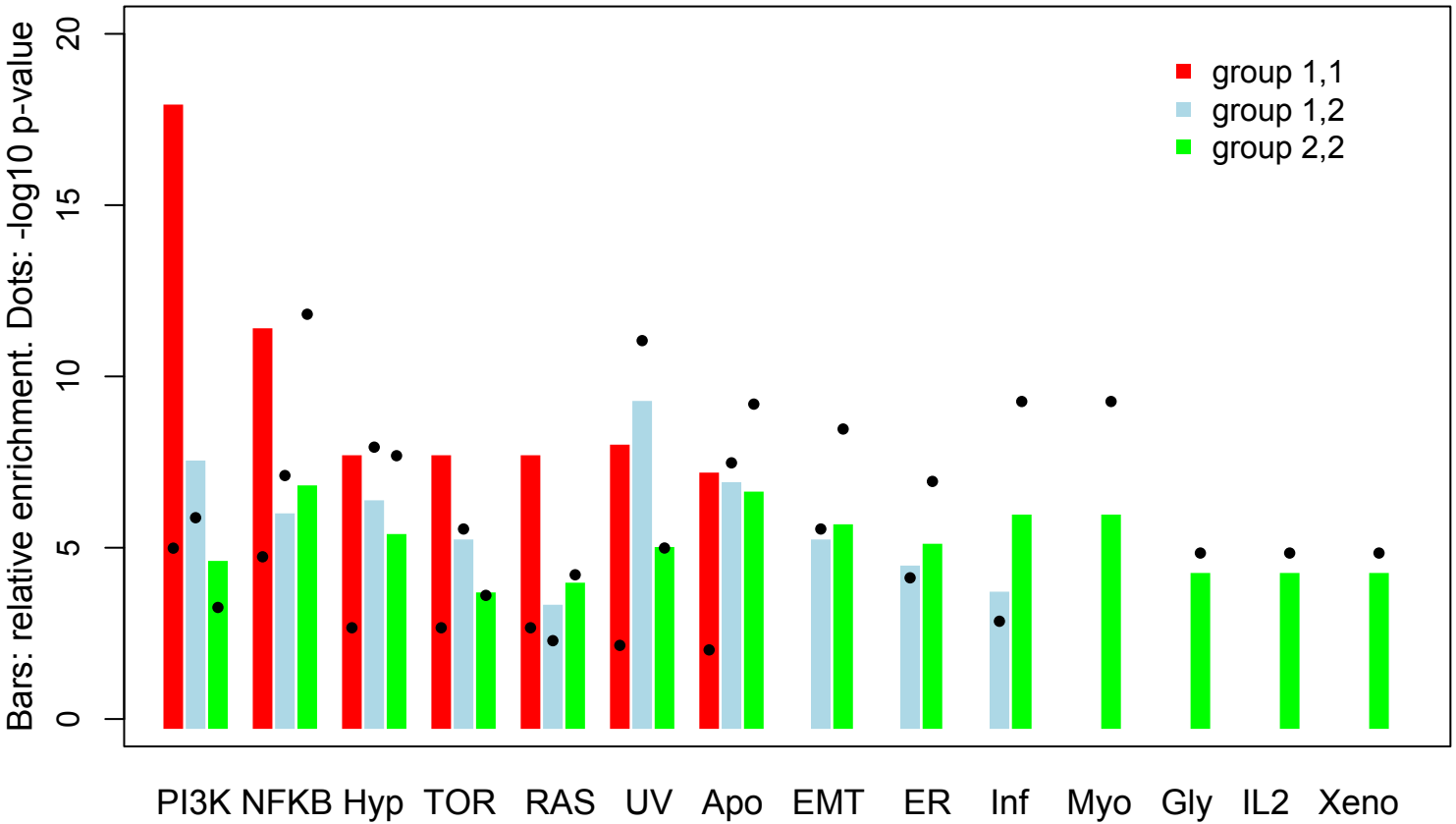
(C) p53 (Transfac)

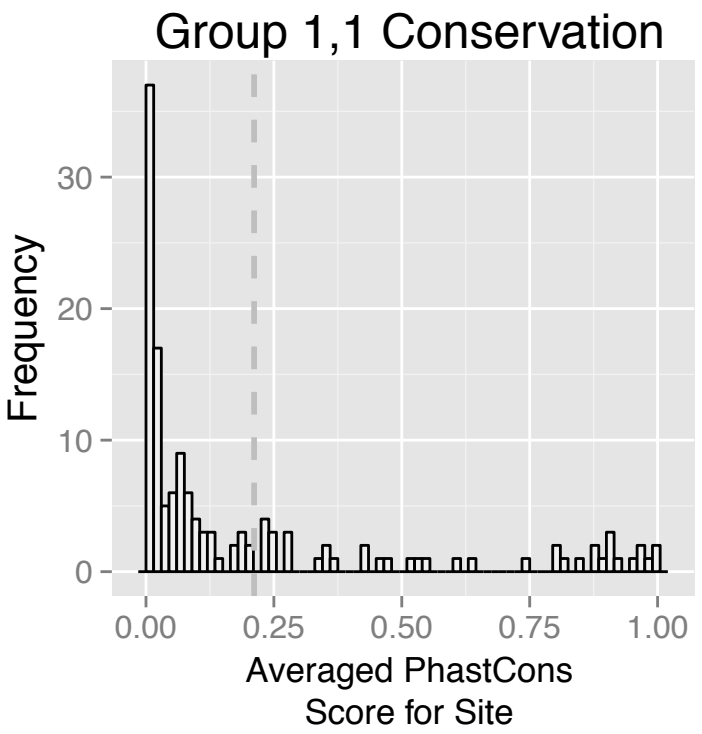


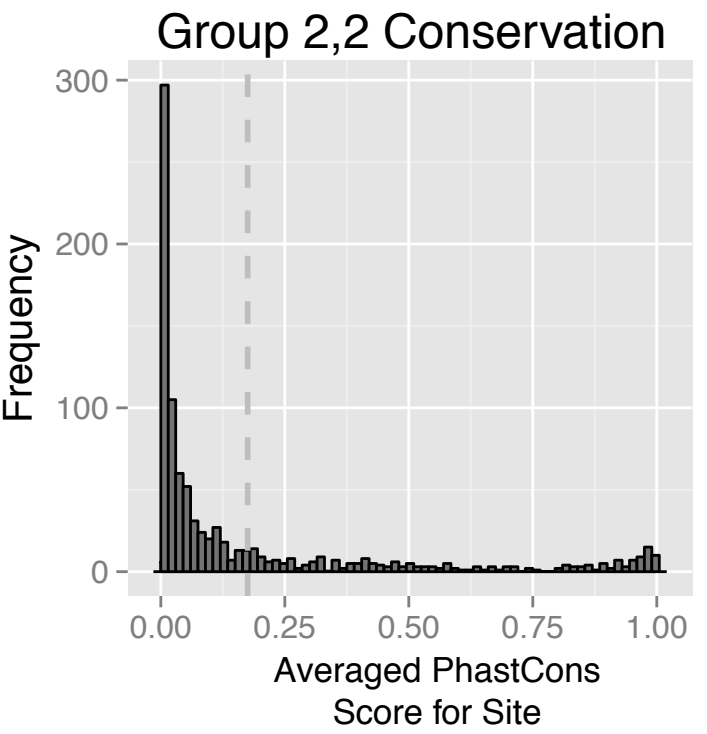
(D) p63

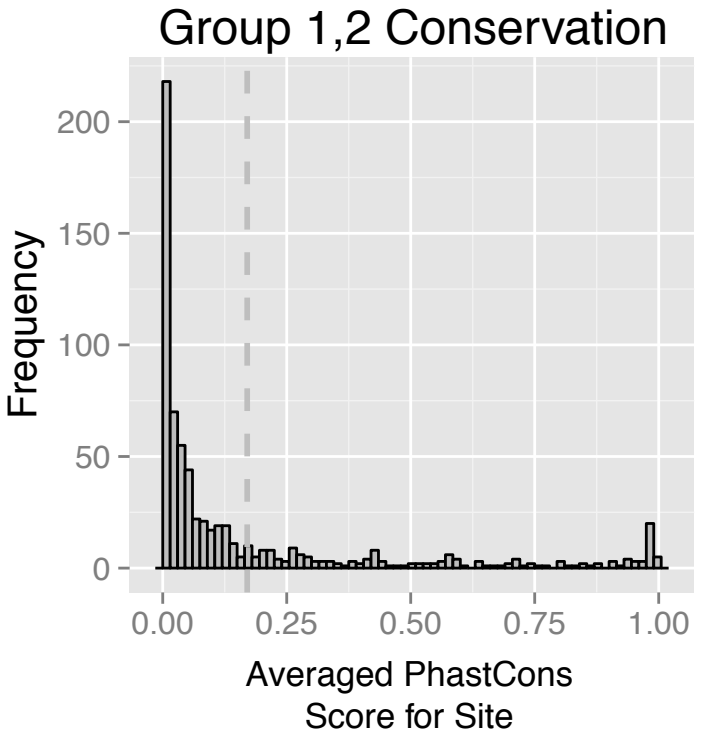


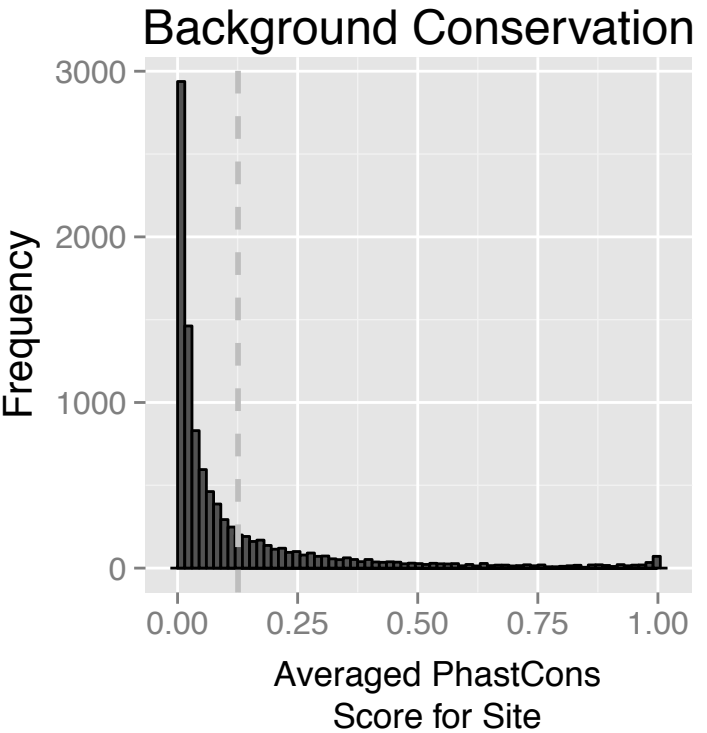
(E) p73


Bars: relative enrichment. Dots: -log10 p-value
0
5
10
15
20
group 1,1
group 1,2
group 2,2
PI3K
NFKB
Hyp
TOR
RAS
UV
Apo
EMT
ER
Inf
Myo
Gly
IL2
Xeno

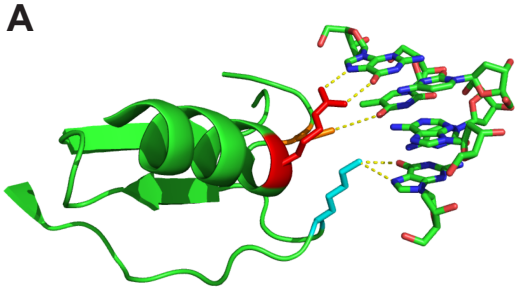

Group 1,1 Conservation
Frequency
0
10
20
30
0.00
0.25
0.50
0.75
1.00
Averaged PhastCons
Score for Site


Group 2,2 Conservation
Frequency
0
100
200
300
0.00
0.25
0.50
0.75
1.00
Averaged PhastCons
Score for Site

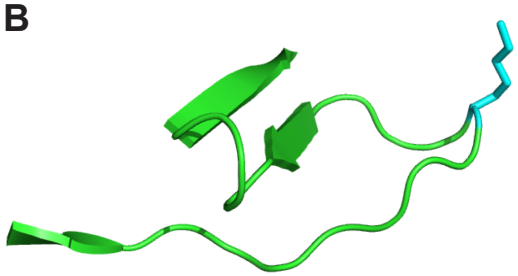

Group 1,2 Conservation
Frequency
0
50
100
150
200
0.00
0.25
0.50
0.75
1.00
Averaged PhastCons
Score for Site


Background Conservation
Frequency
0
1000
2000
3000
0.00
0.25
0.50
0.75
1.00
Averaged PhastCons
Score for Site

A



B



C