Distinct p53 Genomic Binding Patterns in Normal and Cancer-derived Human Cells

Krassimira Botcheva[1], Sean R. McCorkle[1], W. R. McCombie[2], John J. Dunn[1], and Carl W. Anderson[1]

[1]Biology Department, Brookhaven National Laboratory, Upton, NY 11973 USA
[2]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724 USA

SUPPLEMENTAL METHODS:

Cell cultures and treatments

Normal human lung fibroblasts IMR90 were obtained from ATCC (CCL-186, PDL 24) and were grown in EME Medium supplemented with 10% fetal bovine serum, 2 mM L-glutamine, 100 U penicillin and 100 µg streptomycin. Cultures were grown at 37^{o}C in a humidified atmosphere containing 5% CO_2. Cultures were treated with 5-fluorouracil (5-FU, SIGMA) for 6 hrs at a final concentration of 375 µM. Data were generated using cells at PDL 38-40.

Chromatin Immunoprecipitation

ChIP experiments were performed according to Millipore (formerly Upstate Biologicals, cat# MCPROTO407) with some modifications. Briefly, cells were grown to 80 % confluency; after treatment for 6 hrs with 375 µM 5-FU, they were fixed with 1% formaldehyde for 10 min, at room temperature, harvested in PBS supplemented with protease inhibitors (Complete EDTA-free Protease Inhibitor Cocktail; Calbiochem III, cat# 539134) and lysed. The chromatin was sonicated on ice to an average size of 200-500bp using a cup horn Sonicator 3000 (Misonix), at power level 9 (10 pulses of 30 seconds with 1 min breaks). At that point the Input chromatin sample was taken, kept at 4^{o}C and then treated as the ChIP samples. For each ChIP sample, 14 µl of Dynabeads Protein G (Invitrogen, cat# 100.03D) were conjugated with 5µg mouse monoclonal p53-specific antibody DO1 (Santa Cruz, cat #SC-125X) or 5µg non-immune mouse normal IgG (Santa Cruz, cat# sc-2025) and incubated with the soluble chromatin overnight at 4°C. Antibody chromatin complexes were purified, and cross-links were reversed by incubation at 65°C for 4h. All samples (ChIP and Input) were then incubated with RNaseA (Roche), followed by incubation with Proteinase K (Invitrogen) and phenol/chloroform (Sigma) extraction. The DNA was precipitated with ethanol in presence of Glycoblue (Ambion, cat# AM9516), collected, washed three times with 80 % ethanol and then cleaned up using the QIAGEN PCR clean up kit.

Library preparation and Illumina sequencing

ChIP and Input DNA libraries for single end sequencing were prepared with the ChIP-Seq DNA Sample Prep kit from Illumina (cat# IP-102-1001) according to the manufacturer's instructions (Illumina, cat# 11257047) with some modifications based on published improvements (Quail et al. 2008). Briefly, p53 ChIP DNA prepared from 1.5x106 IMR90 cells was purified using the QIAGEN PCR clean up kit, and the fragments were 3' A-tailed. The DNA was purified using the QIAGEN PCR clean up kit with mini-elution columns, and adapters were ligated to the fragments. Adapter-ligated DNA was purified using magnetic beads (Agencourt AMPure kit, cat# A50850), and the entire amount of eluted DNA, without size selection, was used for PCR amplification (14-16 cycles) with Illumina PCR primers 1.1 and 1.2. The PCR cycle number we

determined based on prior analysis of the conditions required to amplify a ChIP DNA sample while keeping at minimum the amplification from an IgG library (negative control library prepared in parallel with the ChIP library). The Input library was prepared from 30 ng Input chromatin.

The PCR amplified ChIP and Input libraries were subjected to electrophoresis on separate gels; fragments with average size ~200-400 bp were recovered using the QIAGEN gel extraction kit, followed by precipitation with ethanol and three times wash with 80% ethanol. Before high-throughput sequencing, a fraction of each amplified library was cloned into the pCR4-TOPO vector (Invitrogen, cat# K4580-01) for a small scale Sanger sequencing to confirm that the libraries contain the Illumina adapters and were not contaminated with foreign DNA. Prior to single-end Illumina sequencing, each DNA library was quantified on the Agilent Technologies 2100 Bioanalyzer using the High Sensitivity DNA kit (cat# 5067-4626, Agilent Technologies) and single end sequencing was performed on Illumina Genome Analyzer 2x instrument. Each of the libraries was loaded on one lane, and data were collected from one sequencing run.

qPCR

Quantitative PCR (qPCR) was carried out using a Rotor-Gene 3000 from Corbett Research. SYBR Green PCR Master Mix was purchased from Applied Biosystem (cat #4309155), primers targeting specific locations were designed with Primer3 (Rozen and Skaletsky, 2000), purchased from Integrated DNA Technologies (standard desalting quality) and used at 200-300 nM final concentration in a 20 µl qPCR reaction volume. For target specific qPCR, 2 µl ChIP DNA was used as a template. For all targets, qPCR was done on both p53-specific ChIP and nonimmune IgG ChIP (negative control) DNAs. Enrichment in the ChIP samples at specific targets was calculated as a fraction of the Input (%).

Mapping reads to hg18

For both the ChIP and Input libraries, 36 nt end sequences (reads) from Illumina base calls were sorted and collected into non-redundant reads. The distinct read sequences were mapped to human genome hg18 (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/) (excluding random and hap files) via in-house software (written in C) which finds short sequences using suffix arrays pre constructed from chromosome sequences. Because memory limitations prohibited the use of a suffix array of the entire genome, each chromosome was separately searched for unique matches and genome-wide unique sites were determined by merging the chromosome results (via perl script). Locations and orientations were collected for reads with either an exact 36 nt match or with a single mismatch. Reads mapped to multiple locations were discarded, and redundant reads having exactly the same genomic coordinates were counted once.

Peak finding

To identify ChIP peaks, we drew from a published method (Rozowsky et al., 2009). First, a function indicating the number of fragments covering each nucleotide position (the fragment coverage) was calculated from unique read positions and orientations assuming a mean fragment length of 350 bp (estimated from the Bioanalyzer data). Peaks were identified by contiguous runs of coverage above a threshold. Runs separated by less than 350bp were bundled together.

Processing was done in chromosome segments of size on the order of a megabase. Thresholds were determined for a segment from the number of reads contained in the segment, from the maximum value of a coverage function obtained by randomly distributing the same number of reads throughout the segment. Because of the non-repeatable nature of these Monte

Carlo calculations and the desire for a replicable threshold level which is strictly a function of number of segment reads and segment length, a fitting function relating threshold to number of reads per segment was made to three standard deviations above the means of the maxima collected from 50 Monte Carlo simulations performed for each of an array of values of read counts and segment lengths. Levels of both ChIP and Input background counts varied across the genome on different scale lengths, which raised the question of the appropriate choice of segment size. We considered segment lengths of $5x10^5$, $7.5x10^5$, $1x10^6$ and $1.5x10^6$ bp; to minimize dependence on the segment length choice, we considered only those peaks which were identified using all four values.

Rejection of ChIP-seq peaks based on Input-seq data

To select a high-confidence set of enriched locations, the identified ChIP-seq peaks were compared to the Input-seq data to ensure statistical enrichment above the background. ChIP-seq read counts in 1 kb windows centered on the peak maxima were compared to Input read counts in the same window. If the Input counts were zero, they were instead estimated from 10 kb windows. Input counts were first normalized by the proportion to total ChIP-seq reads vs Input reads in the segment. Two tests were employed to ascertain if the number of ChIP-seq reads were statistically different from the number of Input reads in the 1 kb windows. The first was a binomial confidence test, as described by Rozowksy et. al. (2009). The second assumed a Poisson distribution for ChIP and normalized Input reads. To be conservative, for the high-confidence set of ChIP-seq peaks, we required confidence above 99 % in both tests and a minimum 3 fold enrichment above the Input background (63 % of the high-confidence peaks had enrichment 10 fold or greater).

Upon manual inspection, some peaks exhibited an unusual structure of multiple spikes of high coincidences of overlapping reads evenly distributed in both orientations. These were assumed to be mapping or reference genome artifacts and were identified by a Pearson correlation R >0.75 between top and bottom strand reads (36 nt) and then removed. Since peak height reflects the likelihood of a given peak to represent true genomic binding, the enriched 1678 ChIP-seq peaks were ordered by peak height and a final cut requiring peak height of at least 10 was applied. The resultant 743 ChIP-seq peaks, which are statistically significantly enriched above the Input background, constitute the final high-confidence set. Confidence and correlations were calculated using R; additional filtering was done with SQL.

Genomic Correlations

Comparisons with RefSeq genes, CGIs, and ChIP-PET (Wei et al., 2006) were performed by downloading table data from the UCSC Genome FTP server and incorporating these into a local PostgreSQL relational database to allow analyses via SQL queries. Table data from other works cited here were similarly collected from the publication's supplemental materials. Peak maximum position was used for determining proximity between IMR90 ChIP / Input-seq peaks and the variety of genomic features (see below).

p53MH sites

A local copy of the p53MH program (Hoh et al. 2002) was used to predict p53 binding sites (score 75 % and above) in 2 kb intervals centered on the peak maxima (Fig. 3A). Since strong enrichment of p53MH sites was observed within 50 nt +/- of the ChIP-seq peak maxima (see Fig. 3A), only those p53MH sites found within +/- 50 of the peak maxima were reported in Supplemental Tables S2, S7, S8 and S9.

TSS and Refseq genes

ChIP-seq peaks proximity to TSSs shown in Fig. 3B, was analyzed by examining 10kb intervals centered at the peak maxima for annotated TSSs (UCSC). All TSSs found were plotted as a function of the distance to the peak maxima. All RefSeq genes for which the peak maxima occurred between the transcript start and stop, or within 20 kb of either end of the gene, are listed in Table S2. For the detailed comparison between ChIP-seq and Input-seq peaks distributions shown in Fig. 4, see Tables S6A and S6B.

Proximity to TSS reported in Fig. 5 for the 4 data sets (ChIP-seq, IMR90; ChIP-PET, HCT116; ChIP-chip, U2OS and ChIP-seq, U2OS) was determined based on the peak maxima (IMR90) or center of the sites reported (HCT116 and U2OS). Since these studies used different fragments sizes and the reported locations varied by width, we also considered the entire reported intervals, counting TSS distance from the start / stop coordinates for the binding locations. There were no significant differences from the results shown in Fig.5, e.g. 38.9 % of the ChIP-seq peaks (IMR90), 6.5 % of the PET3+ clusters (HCT116), 4.8 % of the ChIP-chip sites (U2OS) and 5.7 % of the ChIP-seq sites (U2OS), were found within +/-1kb of a TSS.

CGIs

ChIP-seq peaks were considered to be associated with CGIs if peak maxima were located within CGI bounds extended by 350nt (the average fragment length for our study). Similarly, HCT116 PET3+ clusters (Wei et al., 2006), U2OS ChIP-chip sites (Smeenk et al., 2008) and U2OS ChIP-seq sites (Smeenk et al., 2011) were considered to be associated with CGIs, if the center of the cluster / site was located within CGI bounds extended by 350nt (Fig.5). Because the binding locations reported by the four studies varied by width, we verified that the results presented in Fig. 5 are not sensitive to the association criteria we applied. We repeated the analysis using CGI bounds extended by 350nt, 175nt and 0nt, and for the four sets of p53 binding sites, we used either a single position (peak maximum for IMR90, center of the site for the others), or the entire interval (between start and stop reported for the binding locations). None of the conditions examined changed significantly the results shown in Fig. 5. The minimum and maximum percentage of sites associated with CGIs were as follows: IMR90 (40.6 - 45.8 %), HCT116 (3.2 - 6.5 %), U2OS ChIP-chip (1.6 - 3.9 %) and U2OS ChIP-seq (3.5 - 5.39 %).

Repeats

Repeats were identified using RepeatMasker at http://repeatmasker.org, (Smit A.F.A., Hubley R. and Green P., published reference not available) and Repbase (Jurka et al., 2005) from the Genetic Information Research Institute (http://www.girinst.org/). ChIP-seq peaks were considered to be in a repeat if the peak maximum occurred inside the repeat bounds.

PET3+ clusters, HCT116

The entire ChIP-PET data set reported by Wei et al (2006) was downloaded from UCSC, and the clusters of overlapping distinct PETs were constructed (see Table S7A). For all analyses and comparisons in this study, we used the PET3+ clusters (composed of 3 or more overlapping PET fragments) since, according to the authors' estimate only ~ 2.3 % of them could result from random sampling. A ChIP-seq peak was considered to overlap with a PET3+ cluster if the peak maximum was located within 350 nt of the cluster bounds (Tables S7B, S8, S9). To verify that the results were not sensitive to the chosen criteria, we counted the overlaps when the peak

maximum was within 175nt and 0nt, respectively, from the PET3+ cluster bounds and obtained exact same numbers. Extending that distance to 1 kb lead to only 1 more overlap detected.

ChIP-chip and ChIP-seq identified p53 binding sites in U2OS

The set of high-confidence 1546 ChIP-chip binding sites reported in U2OS cells (Smeenk et al. 2008) was downloaded from http://nar.oxfordjournals.org/content/36/11/3639/suppl/DC1. The set of 2,132 ChIP-seq binding sites reported in U2OS cells (Smeenk et al. 2011) was downloaded from the NCBI GEO database (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22186; accession number GSE22186). The proximity to TSSs and CGIs (reported in Fig. 5) were determined as explained above.

Reference p53 REs

To report those reference p53 REs (from the list of 168, Table S3) identified by ChIP-seq peaks in IMR90, we compared the genomic coordinates of the peak maxima and the p53 RE center. Only ChIP-seq peaks found within 350 nt (the average genomic fragment length in our study) of a p53RE were considered as peaks that identified the corresponding reference p53 REs. These are reported in Table S4 ordered by offset peak maximum – p53RE center. The offset was less than 100 nt for 40 out of the 48 identified reference p53 REs.

ChIP-seq peaks correlation with CpG methylation

The CpG methylation data generated in IMR90 cells (Lister et al. 2009) were downloaded from http://neomorph.salk.edu/human_methylome/. Methylation density plots in Fig 6 were constructed by collecting ratios of the methylated to total Cs (CpG context) for reported sites which occurred in 10 kb windows centered on peak or CGI centers, which were then averaged in bins of 100-200 nt and plotted as a function of the offset from peak maxima or CGI centers.

Motif analysis

For de novo motif analysis we used MEME 4.6.1 (Bailey and Elkan, 1994) via the MEME-ChIP web service (http://meme.sdsc.edu/) and also by running MEME locally on 350 bp regions centered at the peak maxima. All 743 high-confidence peaks were used without selection of training sets. Putative motifs were evaluated by scanning larger (up to 5 kb) regions surrounding the peaks with the associated program MAST (Bailey and Gribskov, 1998) and examining offset histograms accumulated from this output (Fig.7B). We searched for known transcription factor binding sites using TOMTOM (MEME suite 4.6.1 web services). All retrieved putative motifs were evaluated as explained above to evaluate the enrichment pattern.

Functional annotation analysis

Gene ontology analysis was conducted using the Functional Annotation Chart and Functional Annotation Clustering services of the DAVID version 6.7 (Huang da et al., 2009a and 2009b), by preparing and uploading lists of genes associated with the set of 743 high confidence peaks. A ChIP-seq peak was classified as gene associated if the peak maximum was located inside a gene (exon or intron), or within 20 kb of either end of the gene. According to these criteria, of the 743 high-confidence peaks, 153 were intergenic and 590 were associated with 686 genes. Of these 590 peaks, 309 were located within 2 kb of the corresponding gene TSS, 79 within 2-20kb of the gene TSS, 34 were found within 20kb of gene's 3'end and the remaining 168 resided either in a gene exon (21) or intron (147).

To look for enriched signaling pathways, we used the list of 686 genes, DAVID annotation chart analysis and the Kyoto Encyclopedia of Genes and Genomes (KEGG). Only the highest confidence results are shown on Fig 8A (P-value <0.01, 10 fold more stringent than the default DAVID settings).

For functional annotation of all linked gene ontology (GO) terms, we used the list of 686 genes and DAVID functional clustering, which ranks the overall enrichment of entire groups of enriched GO terms based on functional similarity, thus reducing the report redundancy. A total of 177 gene clusters were identified, and those 87 clusters with enrichment score 0.5 and above, as defined by DAVID, without manual selection, are listed in Table S10, ordered by significance (enrichment score). Based on the enriched terms in each cluster (Table S10), cluster names were assigned and reported in Fig. 8B (most highly enriched clusters with enrichment scores above 1.3) and in Fig. S10 (clusters with enrichment score 0.5-1.3).

SUPPLEMENTAL REFERENCES

Bailey TL, and Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol *2*, 28-36.

Bailey TL and Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches, Bioinformatics, *14*:48-54

Hoh J, Jin S, Parrado T, Edington J, Levine AJ and Ott J (2002) The p53MH algorithm and its application in detecting p53-responsive genes. Proc Natl Acad Sci U S A *99*, 8467-8472.

Huang DW, Sherman BT, Lempicki RA (2009a) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. *4*:44-57.

Huang DW, Sherman BT, Lempicki RA (2009b) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*:1-13.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogentic and Genome Research *110*:462-467

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM*, et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315-322.

Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ (2008) A large genome center's improvements to the Illumina sequencing system. Nat Methods *5*: 1005-1010.

Rozen S and Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) *Bioinformatics Methods and Protocols: Methods in Molecular Biology.* Humana Press, Totowa, NJ, p.365-386

Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, and Gerstein MB (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol *27*, 66-75.

Smeenk L, van Heeringen SJ, Koeppel M, van Driel MA, Bartels SJ, Akkers RC, et al. (2008) Characterization of genome-wide p53-binding sites upon stress response. Nucleic Acids Res, *36*:3639-54.

Smeenk L, van Heeringen SJ, Koeppel M, Gilbert B, Janssen-Megens E, Stunnenberg HG, et al. (2011) Role of p53 serine 46 in p53 target gene regulation. PLoS One. , *6*: e17574.

Wei C-L, Wu Q, Vega VB, Chiu KP, Ng P, Zhang T, et al. (2006) A global map of p53 transcription-factor binding sites in the human genome. Cell 124:207-19.

SUPPLEMENTAL FIGURES

Figure S1. Verifying p53 induction and enrichment prior to Illumina library construction.

Figure S2. p53 ChIP-seq overview.

Figure S3. Examples of high-confidence ChIP-seq peaks identified at known reference p53 REs or in their vicinity.

Figure S4. qPCR validation of high-confidence ChIP-seq peaks found in the vicinity of reference p53REs.

Figure S5. Examples of high confidence ChIP-seq peaks, randomly selected for validation among peaks with height below 25.

Figure S6. p53 ChIP-seq peaks identified at bidirectional promoters.

Figure S7. Correlation between detected ChIP-seq peak height and experimentally obtained qPCR enrichment.

Figure S8. Distribution of p53 binding sites with respect to CGIs and TSSs.

Figure S9. Distribution of human CGIs and overlapping p53 ChIP-seq peaks with respect to the nearest TSS.

Figure S10. DAVID functional clustering of the genes associated with high confidence p53 ChIP-seq peaks in IMR90.

SUPPLEMENTAL TABLES

Table S1. Illumina sequencing runs statistics.

Table S2. List of 743 high-confidence p53 ChIP-seq peaks identified in IMR90 cells.

Table S3. Reference list of 168 previously reported functional p53 REs.

Table S4. Reference p53 REs identified by ChP-seq in IMR90 cells.

Table S5. List of high-confidence ChIP-seq peaks identified at biderectional promoters.

Table S6. Distribution of peaks with respect to RefSeq genes.

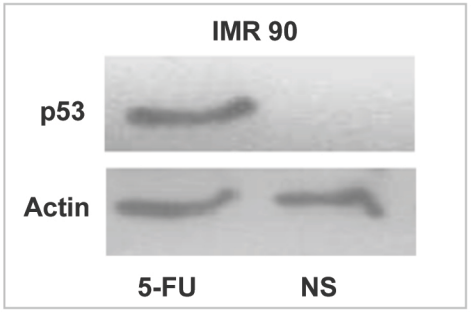
Table S7. Comparison of p53 ChIP-seq peaks identified in IMR90 cells with p53 ChIP-PET clusters reported in HCT116 cells.

Table S8. List of p53 ChIP-seq peaks identified in IMR90 cells overlapping p53 ChIP-PET3+ clusters reported in HCT116.

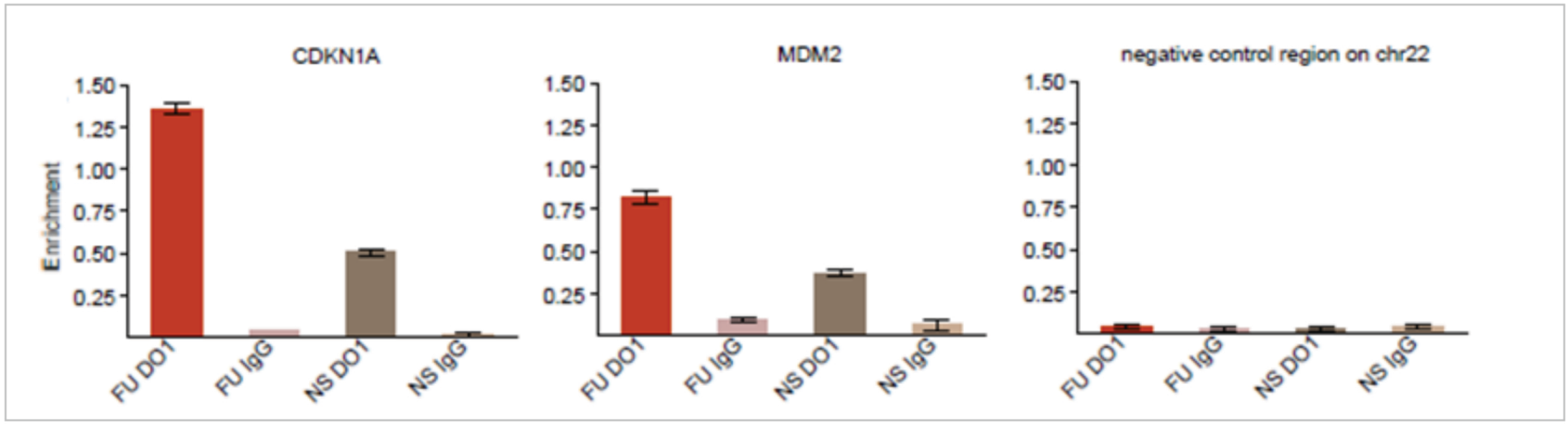
Table S9. Genomic features associated with the p53 binding sites in IMR90 and HCT116 cells.

Table S10. DAVID functional annotation clustering of the genes associated with the 743 high-confidence p53 ChIP-seq peaks in IMR90 cells.

S1A



S1B

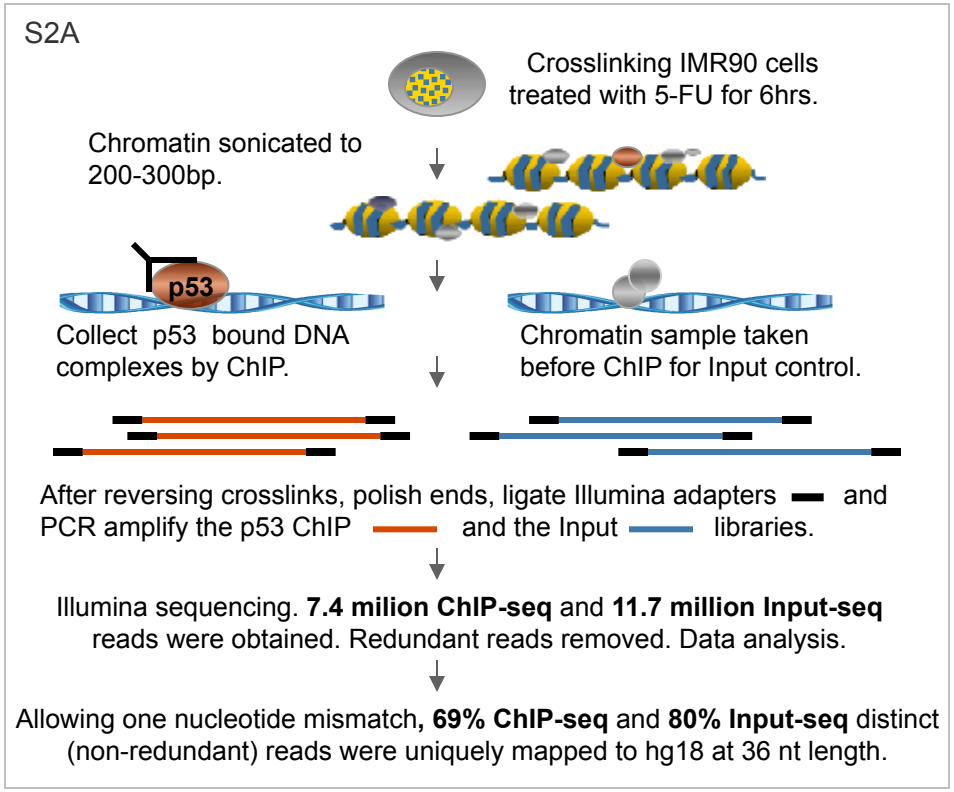


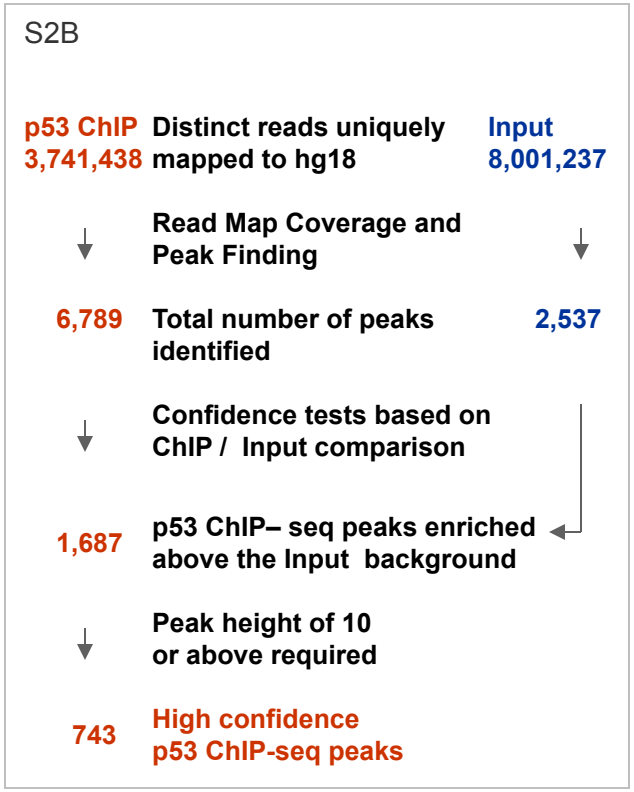
Supplemental Figure S1. Verifying p53 induction and enrichment prior to Illumina library construction.

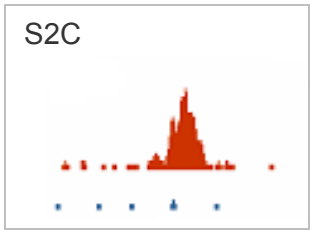
S1A. Whole cell extracts were subjected to Western analysis using the p53-specific antibody DO1. Shown are results from normal IMR90 human fibroblasts after 6 hrs treatment with 375 μM 5-FU. Actin is used as a loading control.

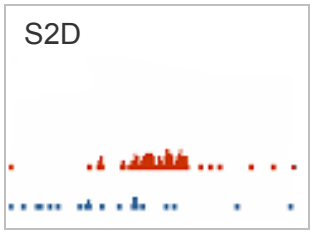
S1B. Target-specific qPCR analysis confirmed p53 enrichment at known binding sites prior to library construction. p53 enrichment is shown at two canonical targets, the *CDKN1A* site at -2,232 bp relative to the TSS (chr6:36,752,204-36,752,224) and at the *MDM2* promoter (chr12: 67,488,970-67,488,990). A region from chromosome 22 (chr22:47,056,575-47,056,854) was used as a negative control for p53 binding. Coordinates are given in hg18.

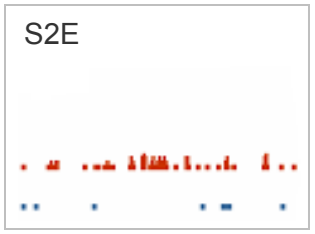
Average enrichment is calculated as percentage of Input; results shown are from duplicated qPCR samples. FU (6hrs treatment with 5-FU); NS (no stimulation); DO1 (ChIP with p53 specific DO1 antibody) and IgG (ChIP with non-specific IgG).

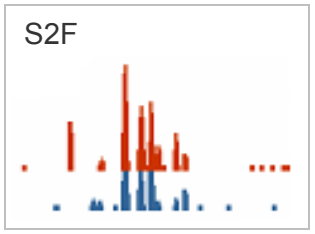


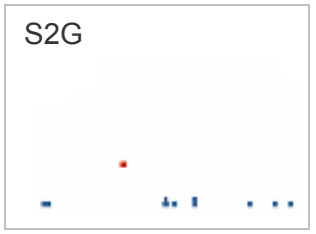






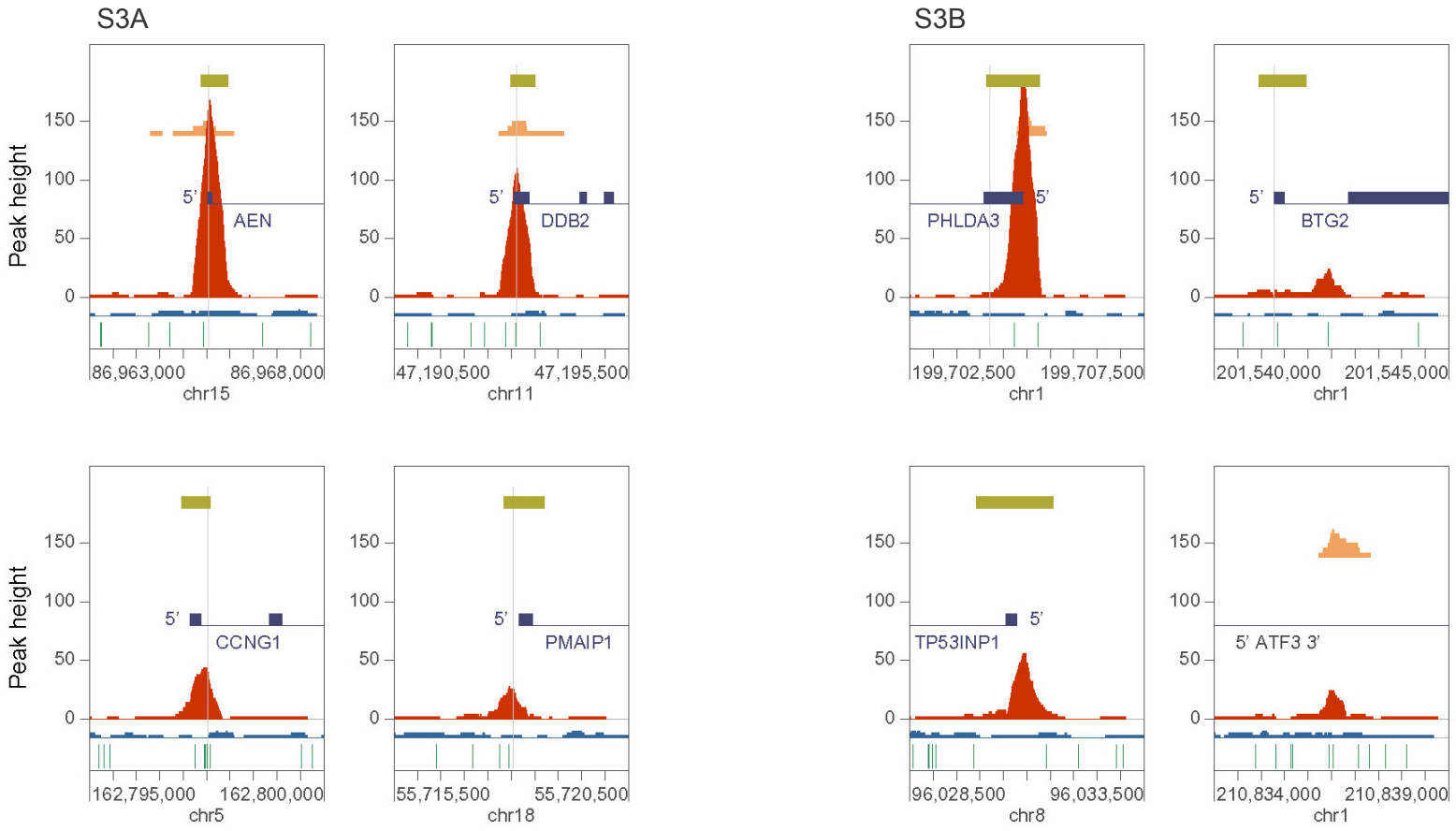






Supplemental Figure S2. p53 ChIP-seq overview.

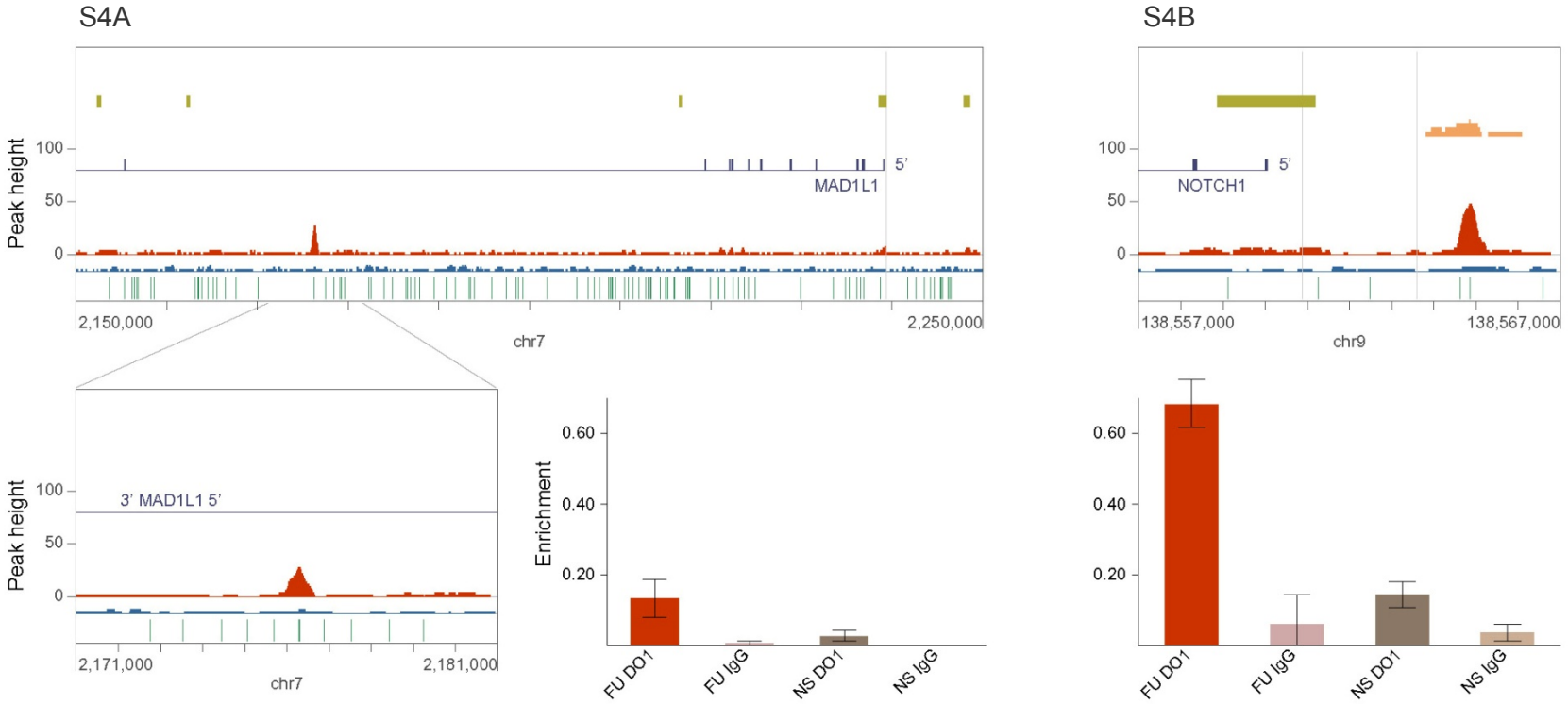
S2A. Experimental procedure. **S2B.** Bioinformatics processing for defining high-confidence p53 ChIP-seq peaks.
S2C-G. Examples of four major profiles of ChIP-seq peaks observed in the study: **S2C.** Sharp ChIP-seq peak of varying height, low Input background (73 % of the high-confidence set); **S2D.** Wide ChIP-seq peak, low Input background (21 % of the high confidence set); **S2E.** ChIP-seq reads covering area of enrichment above the Input background (6 % of the high confidence set); **S2F.** ChIP-seq peaks overlapping Input-seq peaks (excluded from the high confidence set); **S2G.** ChIP-seq map at the chromosome 22 locus, used in this study as a negative control for p53 binding. ChIP-seq reads are plotted in red, Input-seq reads are plotted in blue.



Supplemental Figure S3. Examples of high-confidence ChIP-seq peaks identified at known reference p53 REs or in their vicinity.

S3A. Reference p53 REs matched exactly by high-confidence ChIP-seq peaks at the target genes *AEN*, *DDB2*, *CCNG1* and *PMAIP1*. (see Table S4A for complete list). **S3B.** High-confidence peaks found in the vicinity of reference p53 REs not identified by ChIP-seq, at the target genes: *PHLDA3*, *BTG2*, *TP53INP1* and *ATF3*. Note: In the case of *TP53INP1* and *ATF3*, because of the distance between the ChIP-seq peak maximum and the reference p53RE (respectively 10.7 kb and 11.8 kb), the reference REs fall outside the 5 kb area plotted.

ChIP-seq (red) and Input-seq (blue) coverage maps are plotted in 5 kb regions centered at the peaks. Previously reported functional reference p53 REs are marked with grey vertical lines. All annotated features shown (e.g. RefSeq genes, CGIs, p53 PET3+ and p53MH predicted binding sites) are as indicated on Figure 1.



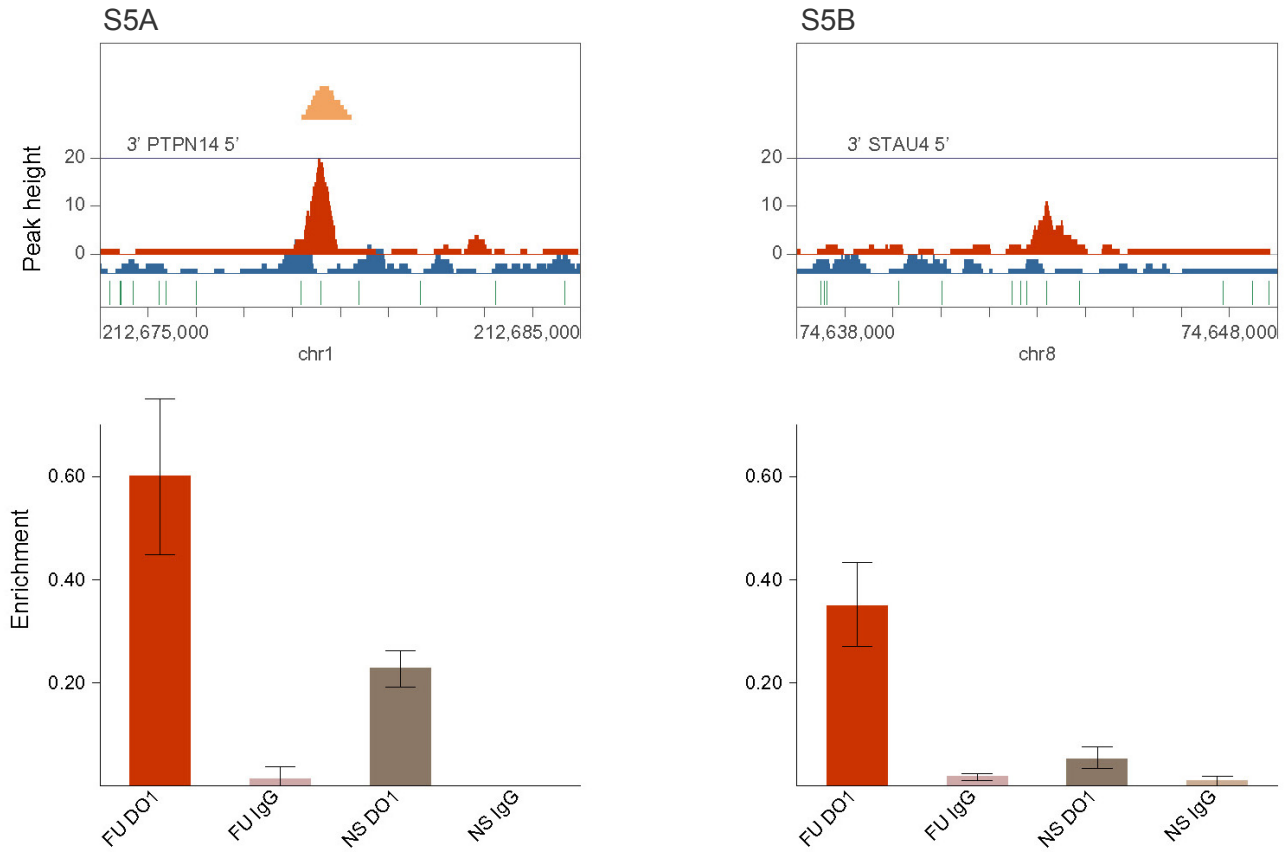
Supplemental Figure S4. qPCR validation of high-confidence ChIP-seq peaks found in the vicinity of reference p53REs.

S4A. qPCR validation of high-confidence peak ID 123 at *MAD1L1*. The reference p53RE at the promoter of *MAD1L1* (see Table S3) was not identified by a high confidence ChIP-seq peak (although a sub-threshold enrichment was detected). The confirmed peak (ID 123, see Table S2) was internal to the gene.

S4B. qPCR validation of high-confidence peak ID 53 at *NOTCH1*. The two reference p53REs (see Table S3) were not identified by ChIP-seq peaks. The qPCR validated high-confidence peak ID 53 (see Table S2) was found in close proximity, intersecting a PET5 cluster.

ChIP-seq (red) and Input-seq (blue) coverage maps are plotted. Previously reported functional reference p53REs are marked with grey vertical lines. All annotated features shown (e.g. RefSeq genes, CGIs, p53 PET3+ and p53MH predicted binding sites) are as indicated on Figure 1.

qPCR validation was done on independently obtained ChIP samples from IMR90, treated for 6hrs with 5-FU. Average enrichment is calculated as percentage of Input; results shown are from duplicated qPCR samples. FU (6hrs treatment with 5-FU); NS (no stimulation); DO1 (ChIP with p53 specific DO1 antibody) and IgG (ChIP with non-specific IgG).

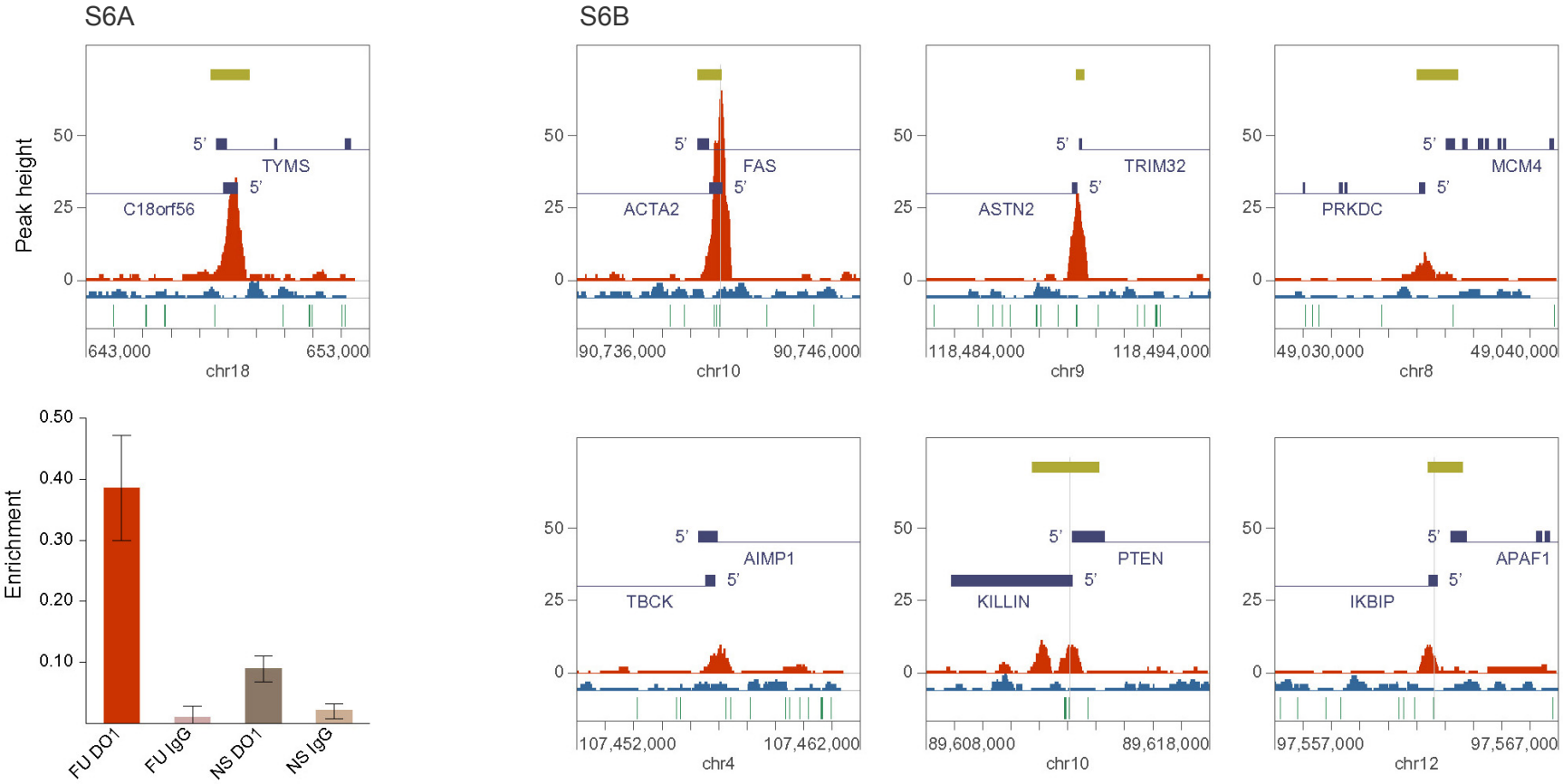


Supplemental Figure S5. Examples of high confidence ChIP-seq peaks, randomly selected for validation among peaks with height below 25.

S5A. qPCR validated peak ID 196 (height 20) at *PTPN14*, intersecting a PET7. **S5B.** qPCR validated peak ID 573 (height 11) at *STAU4*.

ChIP-seq (red) and Input-seq (blue) coverage maps are plotted in 5 kb regions centered at the peaks. All annotated features shown (e.g. RefSeq genes, CGIs, p53 PET3+ and p53MH predicted binding sites) are as indicated on Figure 1.

qPCR validation was done on independently obtained ChIP samples from IMR90, treated for 6hrs with 5-FU. Average enrichment is calculated as percentage of Input; results shown are from duplicated qPCR samples. FU (6hrs treatment with 5-FU); NS (no stimulation); DO1 (ChIP with p53 specific DO1 antibody) and IgG (ChIP with non-specific IgG).

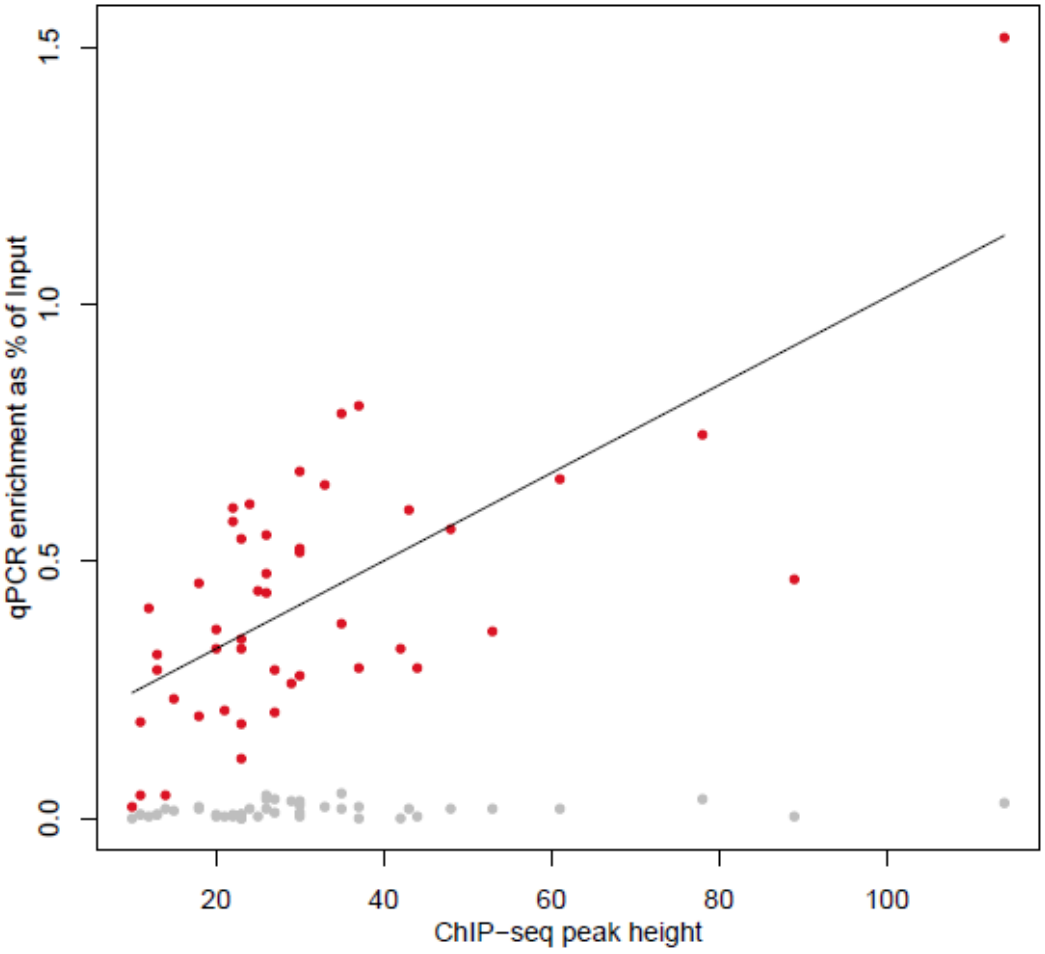


Supplemental Figure S6. p53 ChIP-seq peaks identified at bidirectional promoters.

S6A. qPCR validation of peak ID 84 (Table S2) at the gene pair *TYMS* / *C18orf56*.
S6B. Examples of high-confidence peaks mapped at the gene pairs: *FAS* / *ACTA2*; *TRIM32* / *ASTN2*; *MCM4* / *PRKDC*; *AIMP1* / *TBCK; PTEN* / *KILLIN;* and *APAF1* / *IKBIP*.

ChIP-seq (red) and Input-seq (blue) coverage maps are plotted in 5 kb regions centered at the peaks. Previously reported functional reference p53REs are marked with grey vertical lines (see Table S3). All annotated features shown (e.g. Refseq genes, CGIs, p53 PET3+ clusters and p53MH predicted binding sites) are as indicated on Figure 1.

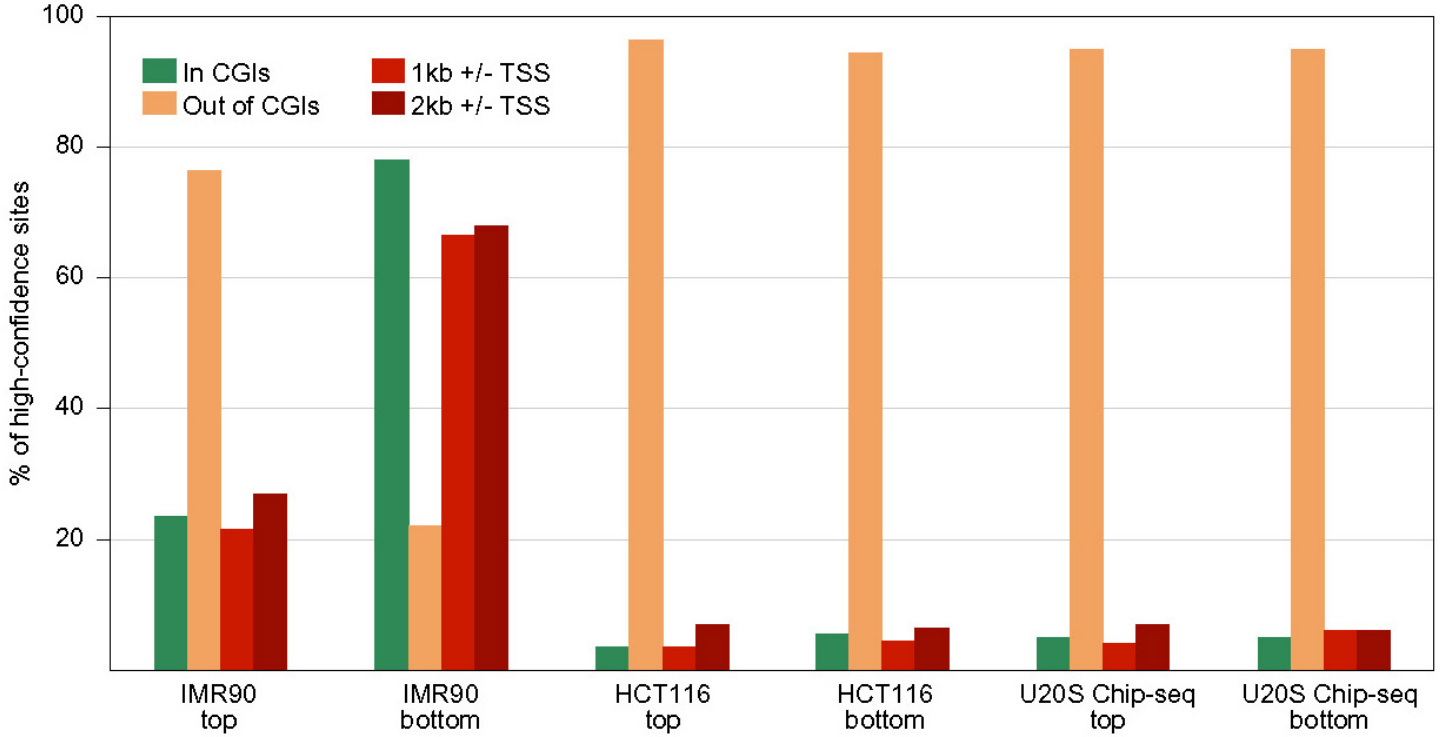
qPCR validation was done on independently obtained ChIP samples from IMR90, treated for 6hrs with 5-FU. Average enrichment is calculated as percentage of Input; results shown are from duplicated qPCR samples. FU (6hrs treatment with 5-FU); NS (no stimulation); DO1 (ChIP with p53 specific DO1 antibody) and IgG (ChIP with non-specific IgG).



Supplemental Figure S7. Correlation between detected ChIP-seq peak height and experimentally obtained qPCR enrichment.

Average qPCR enrichment from two independent ChIP samples is plotted versus peak height for 45 validated high-confidence peaks. Of these, 7 overlapped known reference p53REs, 16 were at locations not reported previously and 22 were randomly selected among peaks with height 25 or below (lower peak height range was specifically targeted for the random validation).

p53-specific enrichment (ChIP DO1, plotted in red) shows a good correlation with the peak height. There is no such correlation between the qPCR enrichment in the non-specific mock control (ChIP IgG, plotted in grey) and the peak height.



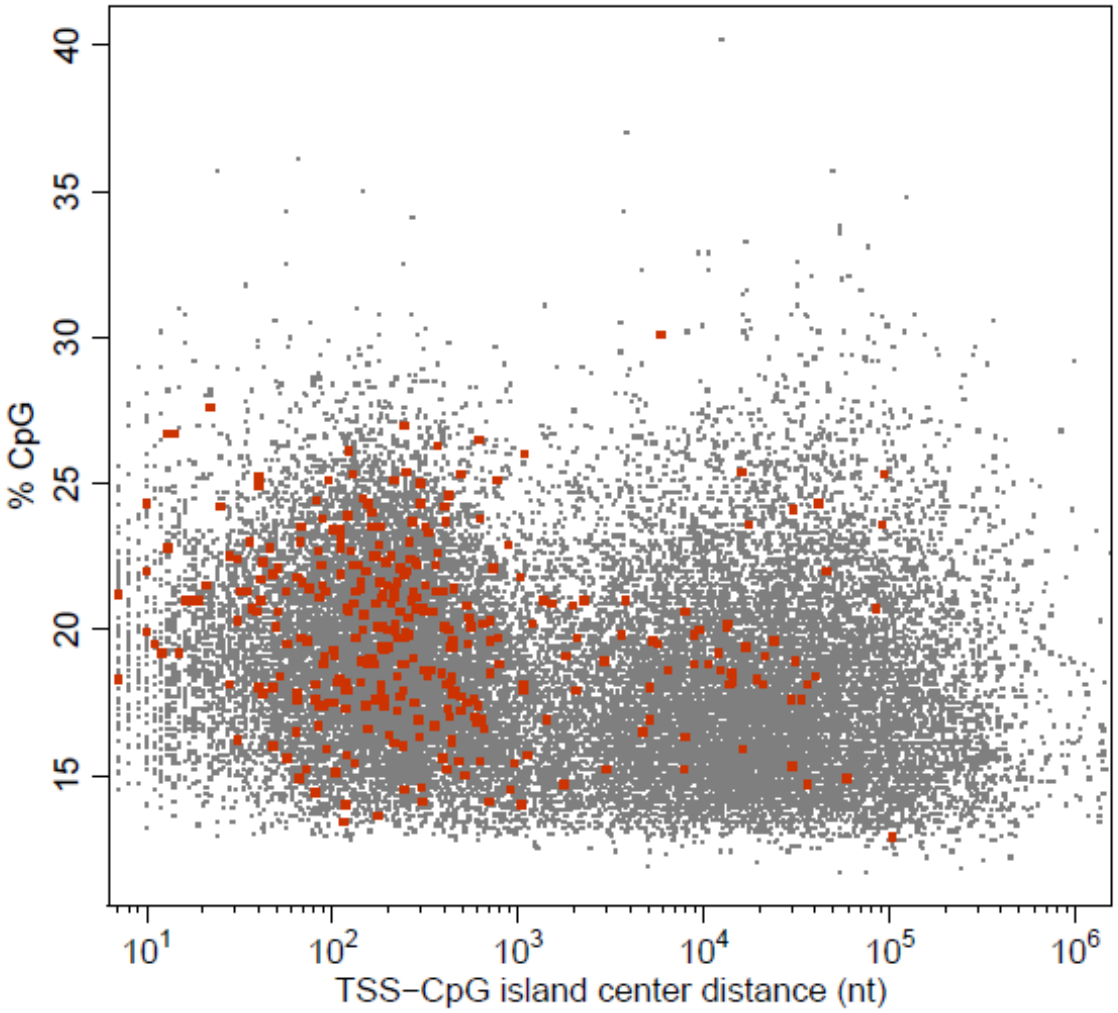
Supplemental Figure S8. Distribution of p53 binding sites with respect to CGIs and TSSs.

In order to evaluate the possibility that the distribution of the p53 binding sites (Fig.5) is strongly affected by the low ranked sites, we compared the top and bottom fractions of the high-confidence sites identified in IMR90 (743 ChIP-seq peaks), HCT116 (310 PET3+ loci) and U2OS (2,132 ChIP-seq peaks). The ChIP-chip study in U2OS, which does not provide ranking information, was omitted from this analysis.

Fractions for the two ChIP-seq studies were determined based on the peak height. Top fractions: IMR90 ChIP-seq peaks with height 20 and above (~27% of the data set) and U2OS ChIP-seq peaks with height 36 and above (~ 27% of the data set). Bottom fractions: IMR90 peaks with height 10 (~20 % of the dataset, the lowest peak height for our set is 10) and U2OS peaks with height 11 and below (~20 % of all sites, the lowest peak height for that set is 8).

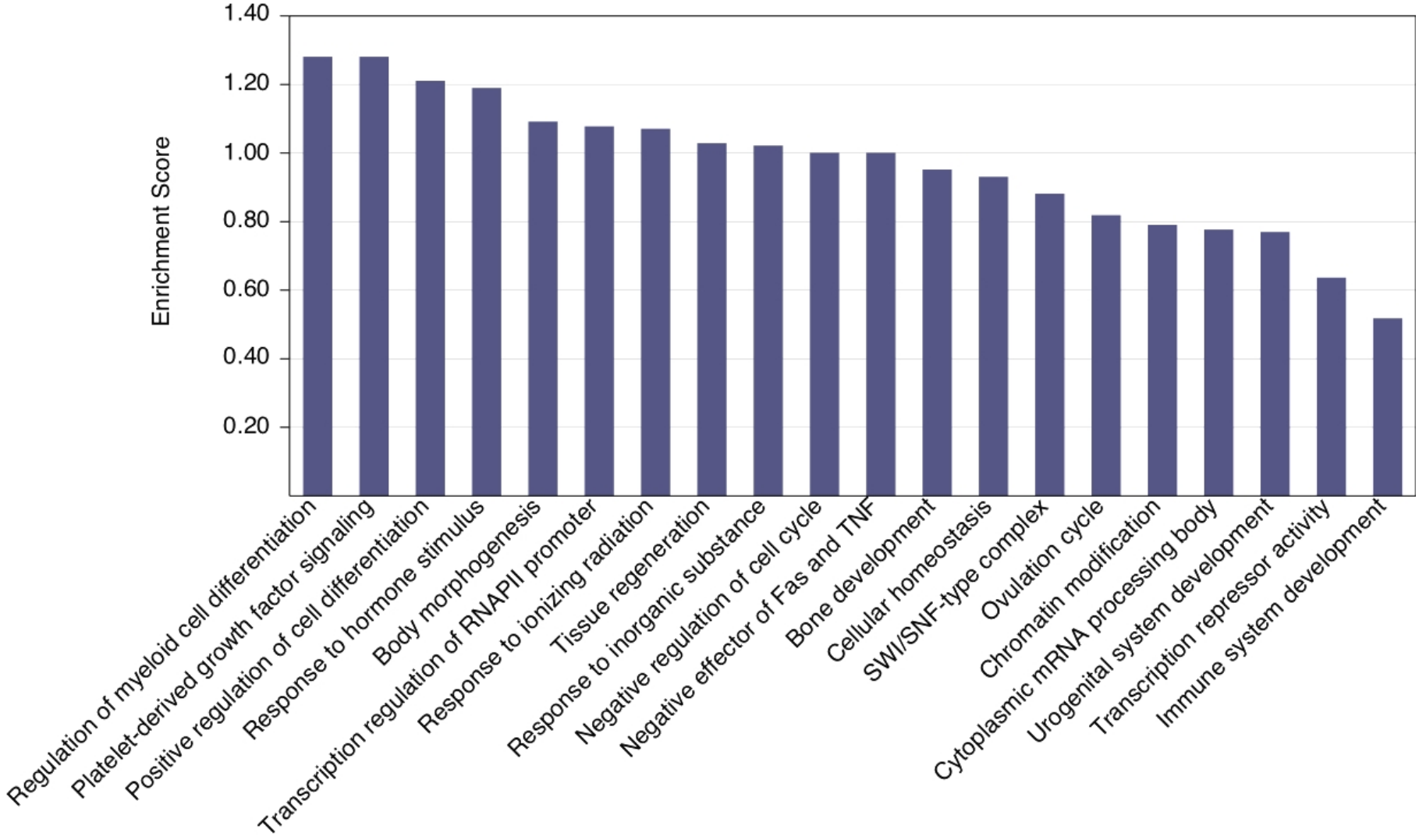
The binding sites annotated by the PET study in HCT116 cells are ranked by cluster number (maximum number of overlapping fragments) . We considered as a bottom fraction all PET3 clusters (157, ~ 50% of all sites) and as a top fraction all PET4+ clusters (153 sites, ~ 49% of all sites).

The top and the bottom fractions of the two cancer datasets show little difference; each of these are equally less enriched at TSS and CpG islands. There is a noticeable difference between the top and bottom fractions of the IMR90 dataset, but despite that, both are clearly different from any of the cancer datasets, showing higher enrichment at TSS and CpG islands.



Supplemental Figure S9. **Distribution of human CGIs and overlapping p53 ChIP-seq peaks with respect to the nearest TSS**.

CpG fraction of CGIs (%) is plotted as a function of distance to the nearest TSS (measured from the CGI center). All 27,639 human CGIs (UCSC defined, hg18) are shown in grey and those CGIs at which high-confidence ChIP-seq peaks are found (331 out of the 743 high-confidence peaks) are shown in red. p53 ChIP-seq peaks are enriched at proximal CGIs (close to TSS) and less represented at distal CGIs (further away from TSS).



Supplemental Figure S10. DAVID functional clustering of the genes associated with high confidence p53 ChIP-seq peaks in IMR90.

Shown are enriched clusters of genes with enrichment score 0.5 - 1.3. See Table S10 for details.