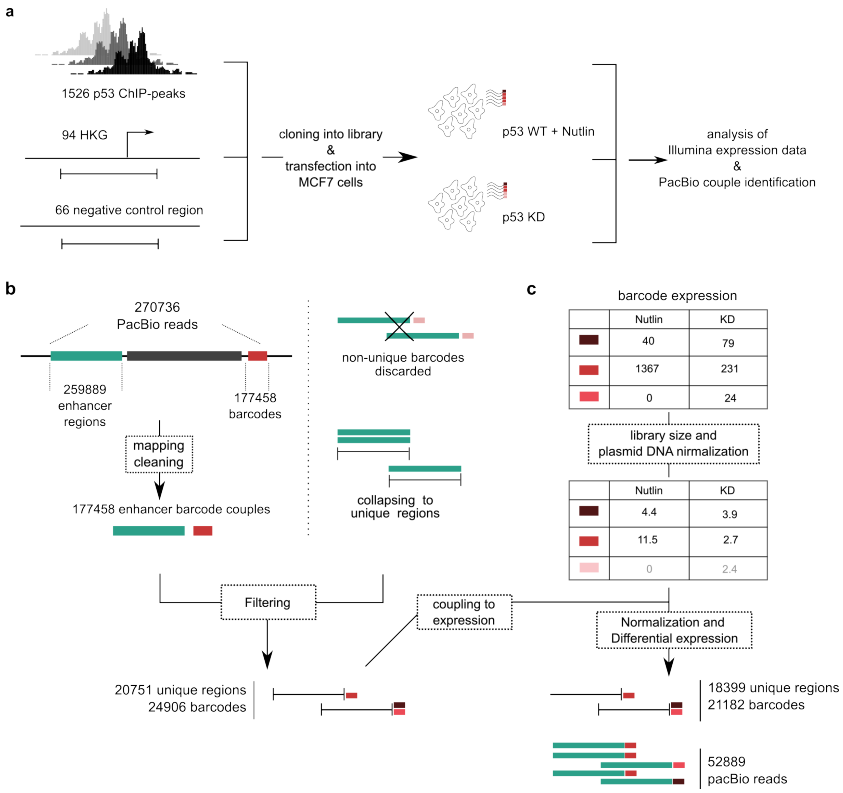
Multiplex enhancer-reporter assays uncover unsophisticated TP53 enhancer logic

Annelien Verfaillie[1], Dmitry Svetlichnyy[1], Hana Imrichova, Kristofer Davie[1], Mark Fiers[2], Zeynep Kalender Atak[1], Gert Hulselmans[1], Hana Imrichova[1], Valerie Christiaens[1], and Stein Aerts[1#]

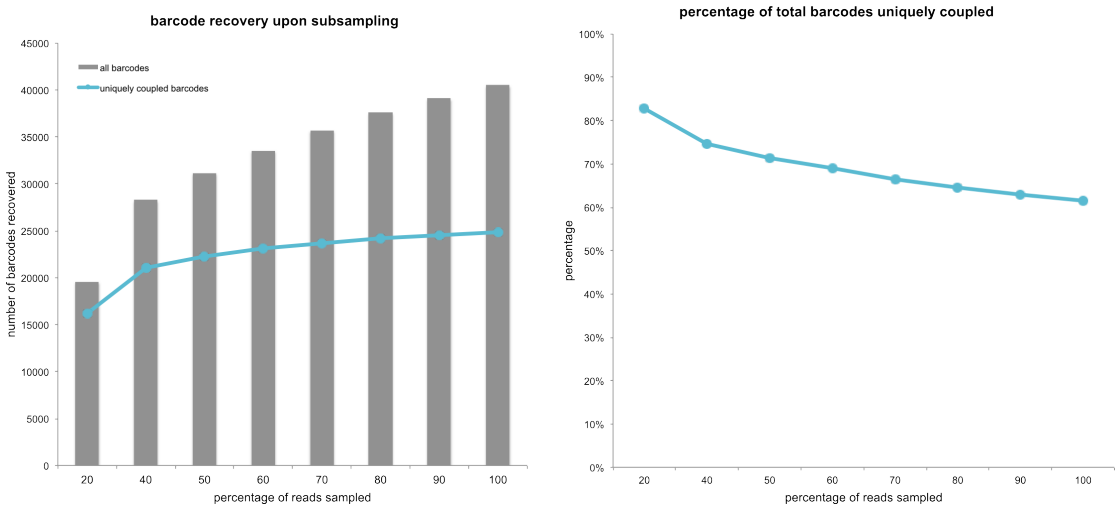
[1]Laboratory of Computational Biology, Center for Human Genetics, University of Leuven, Leuven, Belgium

[2]VIB Center for the Biology of Disease, Leuven, Belgium

correspondence to: stein.aerts@med.kuleuven.be

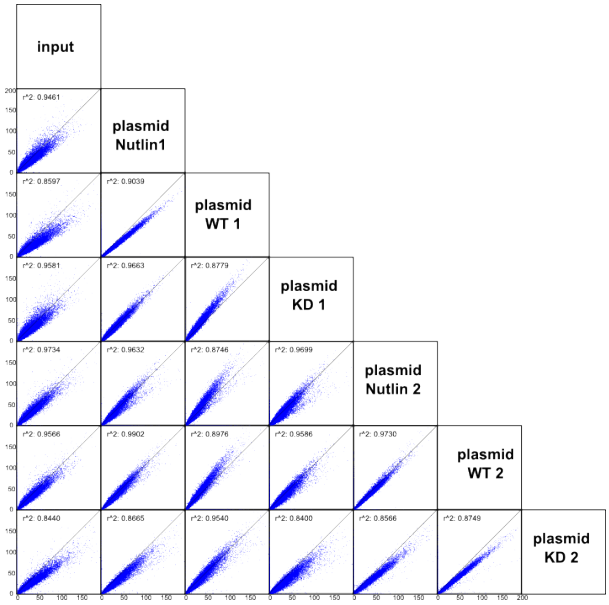


Supplemental Figure S1 - CHEQ-seq applied to a set of candidate TP53 target enhancers. **(a)** TP53 ChIP-peaks in MCF7, determined in a previous experiment were selected based on peak score. Additionally, promoters of housekeeping genes (HKG) and regions with low binding activity in the genome were selected as positive and negative controls respectively. Regions were captured, cloned into the CHEQ-seq plasmid and transfected into MCF7 cells under various conditions (p53WT + Nutlin= TP53 wild type cells stimulated with Nutlin-3a and p53 KD = TP53 knock down). **(b)** Flow chart showing the subsequent steps in PacBio data processing identifying unique reliable region-barcode couples. **(c)** Data processing of the Illumina cDNA barcode expression data and coupling it back to PacBio data to determine the final list of region-barcode couples with corresponding barcode expression data.

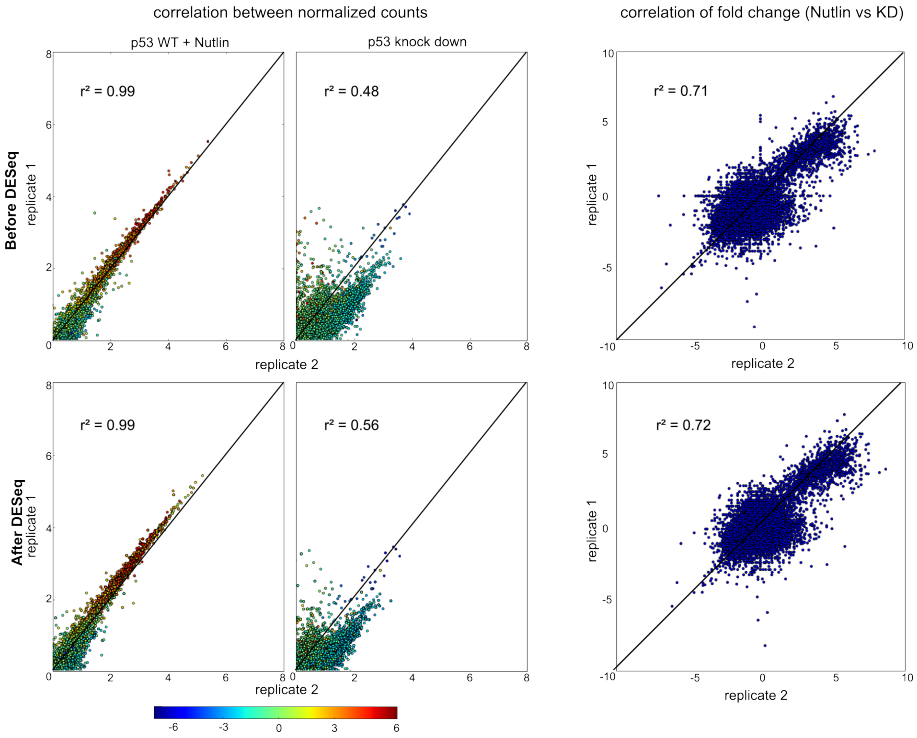


Supplemental Figure S2 – Unique barcode to region coupling.

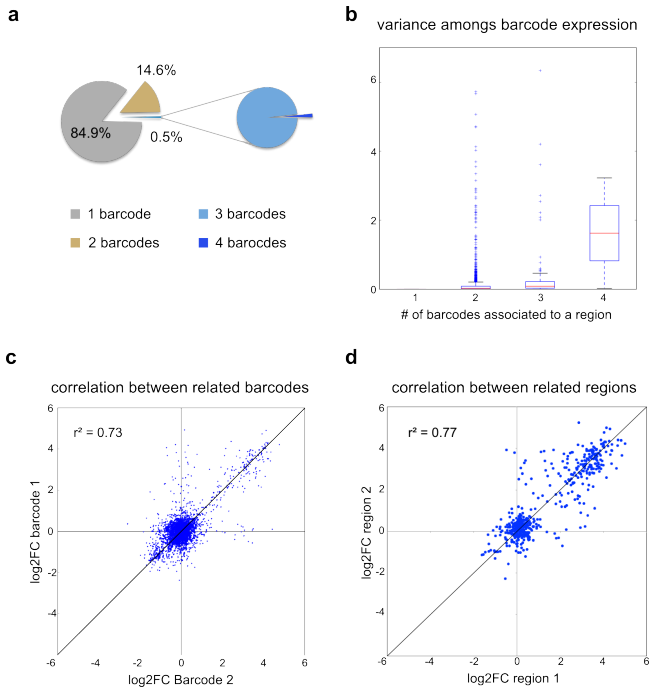
From the 177,458 initial region-barcode couples, we retained 20,751 distinct genomic fragments, linked to 24,906 distinct barcodes, after removing redundancy and retaining only barcodes that pair uniquely with one particular region. The method of assignment of barcodes to regions is robust and additional sequencing would reveal most barcodes to be assigned as unique correctly. The number of all barcodes (grey bars) or uniquely coupled barcode (blue line) that are recovered upon subsampling reads. On the right plot the percentages of barcodes uniquely coupled to a region per subsampling is plotted. The number of uniquely coupled barcodes stagnates towards 100% of reads analysed indicating that more sequencing would not reveal more barcode-regions couples.



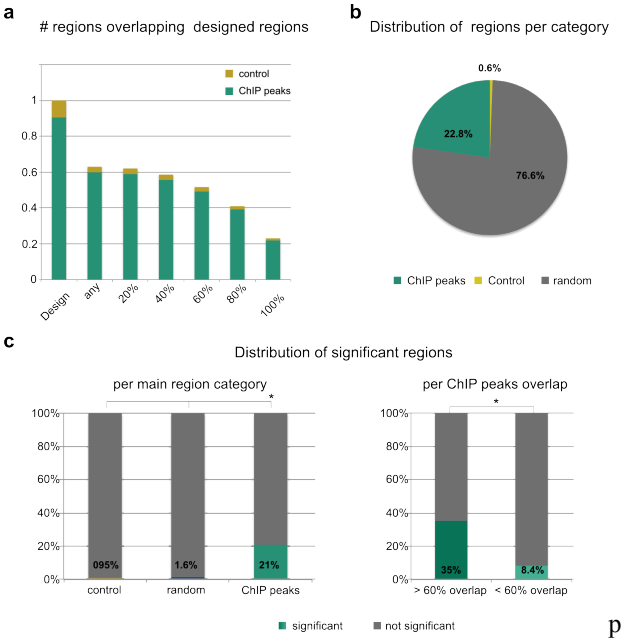
Supplemental Figure S3 - Correlation plot between barcode DNA. Correlation between input and plasmid barcode DNA extracted from cells after transfection. There is strong correlation between all samples, indicating no significant influence of transfection.



Supplemental Figure S4 - Correlation of replicate samples within each condition before and after DESeq normalization. Replicate samples within each condition were correlated to each other in barcode expression levels (p53 WT stimulated with Nutlin (p53 wT +Nutlin) $r^2 = 0.99$ and p53 knock down $r^2 = 0.56$). Differential barcode expression (Nutlin vs KD) correlates strongly between replicate samples ($r^2 = 0.70$). Color scales indicate calculated differential expression levels for each barcode when contrasting Nutlin-3a stimulated samples against TP53 knock down samples (averaging over the replicates).

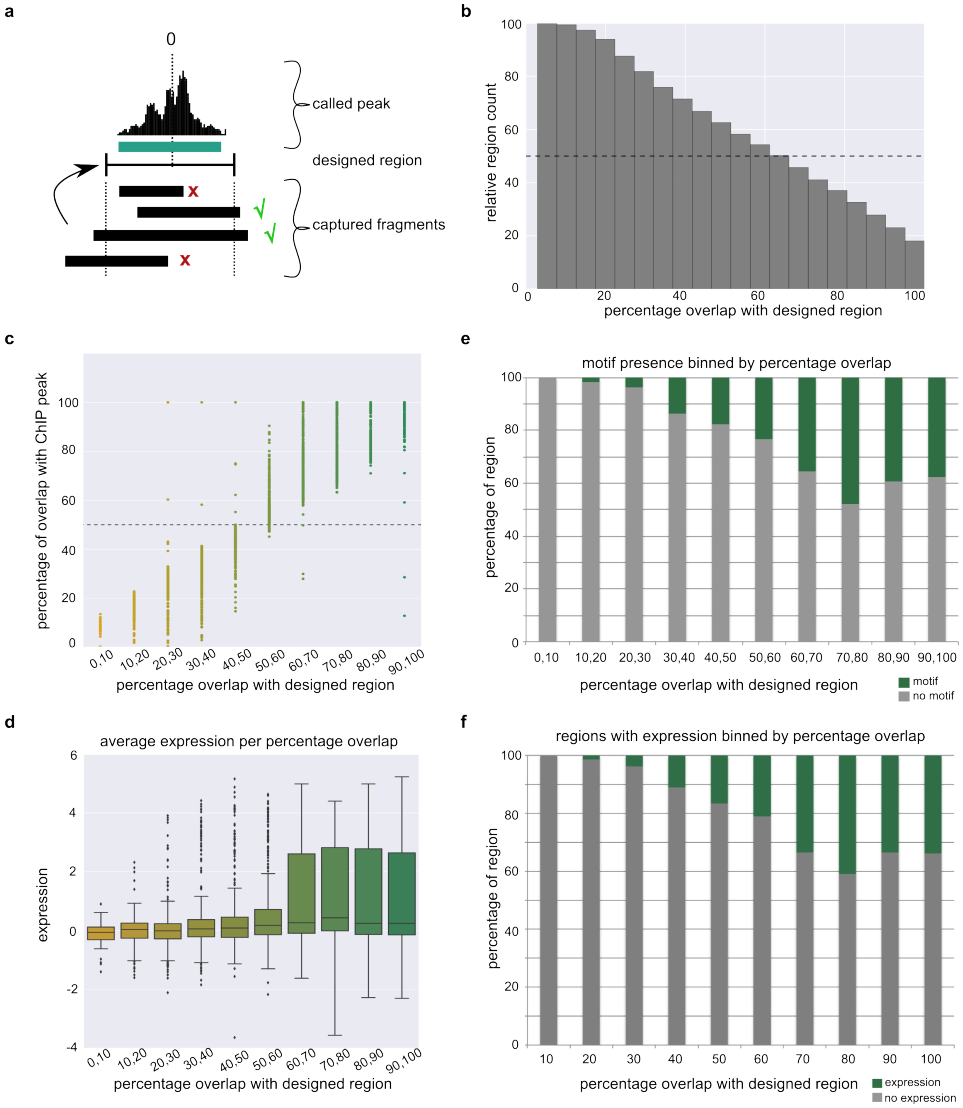


Supplemental Figure S5 - Distribution of barcodes among regions. (a) 15.1% of all regions are represented by more then one barcode uniquely associated to them. **(b)** Boxplot indicating the variance in expression values amongst barcodes assigned to the same region, grouped by the number of barcodes a particular region has. **(c)** The correlation of the expression values between barcodes assigned to the same region, for those regions that have two barcodes assigned to them. **(d)** The correlation between two regions representing the same designed peak. the region with the highest expression value and the region best overlapping it are selected for each designed peak.

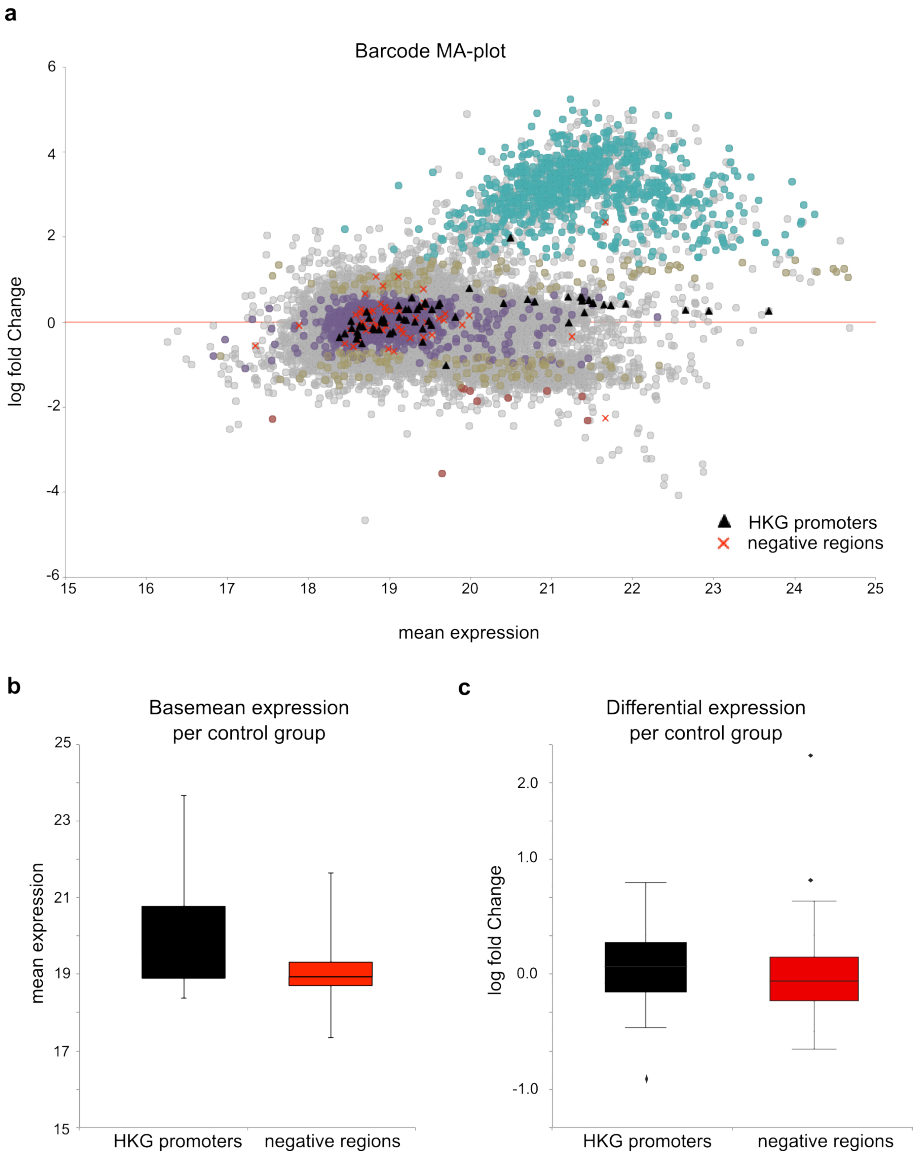


p

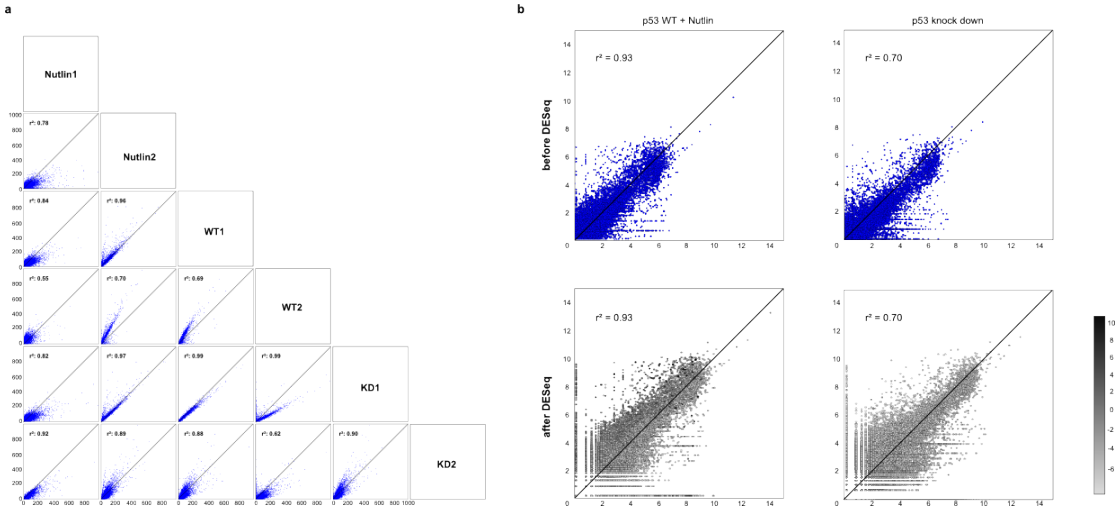
Supplemental Figure S6 – Distribution of CHEQ-seq regions. (**a**) Bar chart showing the gradual decrease in recovered regions upon an increasing overlap with the designed region. **(b)** Distribution of CHEQ-seq regions according to different categories: overlapping with the designed fragments (control or ChIP peaks) or randomly captured from the genome (random) **(c)** Bar charts on the left indicate the relative distribution of significant up-regulated regions (p-val < 0.05, log2FC > 1.5) amongst the different categories of CHEQ-seq regions (* = chi-square test p-val = $1x10^{-274}$). Bar charts on the right represent the distribution of significantly up-regulated regions from the ChIP peak category specifically, when considering an overlap with a ChIP peak of more or less than 60% (* = chi-square test pval = $7.3x10^{-102}$).



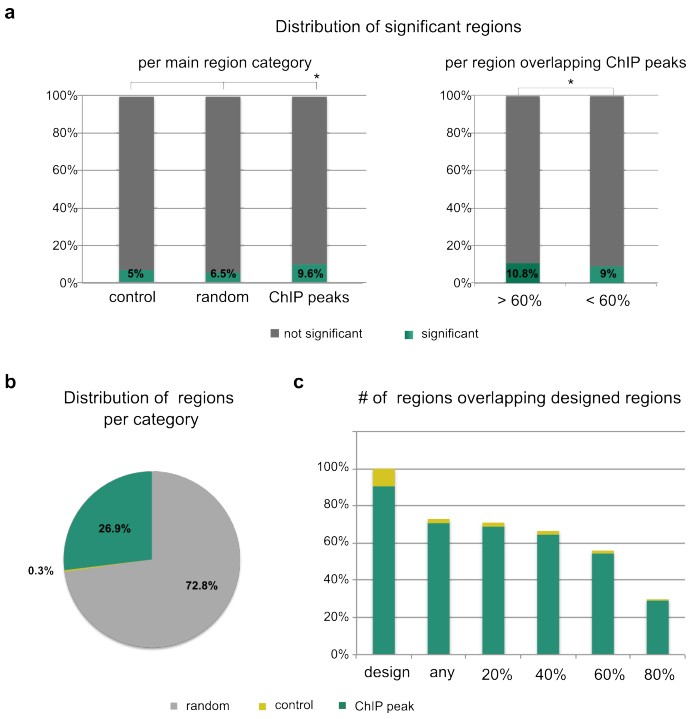
Supplemental Figure S7 – Exploration of the percentage overlap between design and captured fragments. **(a)** Overview of the method of overlapping. Regions were designed based on called peaks. Called peaks were assumed to center around a potential binding site. At least 500bp around the center of the peaks was used as basis for the designed regions. These regions were the targets for the capture**.** Actual captured fragments were mapped against the design and only regions that overlap for 60% with the designed region are retained. **(b)** The number of captured fragments recovered per percentage of overlap with their corresponding design shows that 50% of captured fragments are kept when requiring 60% overlap with the design. **(c)** When overlapping at least 60% with their corresponding design, most regions also overlap at least 50% with the original called peak. **(d)** The average expression coupled to the regions rises markedly above 0 when requiring a minimum overlap of 60% of the captured fragments with the designs. **(e-f)** The percentage of regions with a motif (e) or having a significant expression (log2 fold change >1.5) (f) binned per percentage overlap with the design.



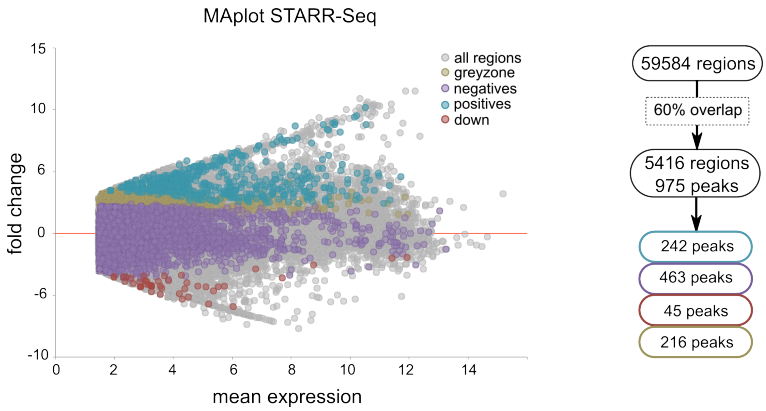
Supplemental Figure S8 – Behavior of control regions in CHEQ-seq. Promoters of housekeeping genes (HKG) (black triangle) and negative regions (red cross) both show no differential expression upon TP53 stimulation **(a, c).** Negative regions have a low basemean expression, while some HKG promoter have a higher basemean expression, as expected. This indicates the ability of the latter to drive expression independently of TP53 **(a-b).**



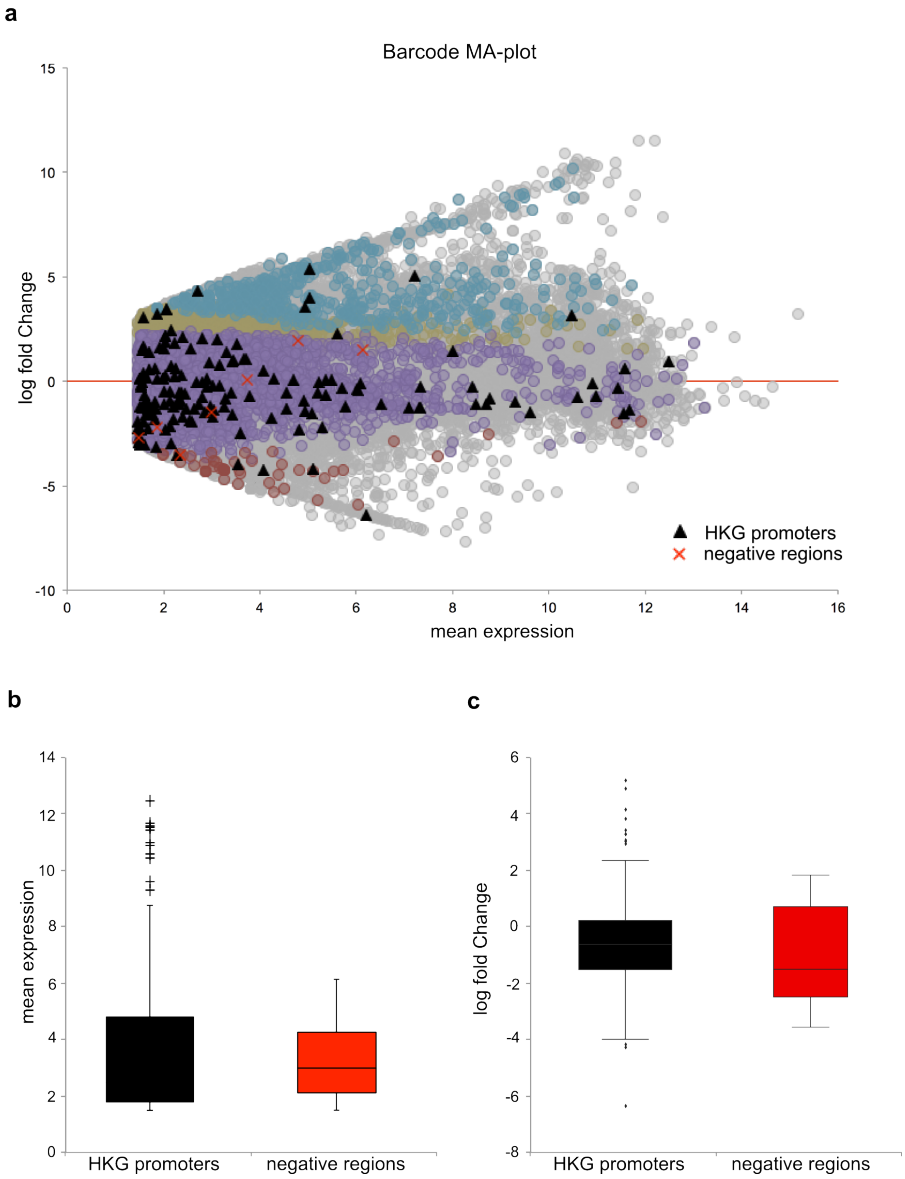
Supplemental Figure S9 – reproducibility of the STARR-seq data. (a) Correlation plot between barcode DNA. Correlation between plasmid barcode DNA extracted from cells after transfection. There is strong correlation between all samples, indicating no significant influence of transfection. **(b)** Correlation of replicate samples within each condition before and after DESeq normalization. Replicate samples within each condition were correlated to each other in barcode expression levels (TP53 WT stimulated with Nutlin (p53 WT +Nutlin) $r^2 = 0.93$ and p53 knock down $r^2 = 0.70$). grey scales indicate calculated differential expression levels for each barcode when contrasting Nutlin-3a stimulated samples against TP53 knock down samples (averaging over the replicates).



Supplemental Figure S10 – Distribution of STARR-seq regions. **(a)** Bar charts on the left indicate the relative distribution of significant regions (p-val < 0.05, log2FC > 1.5) amongst the different categories of STARR-seq regions (* = chi-square test p-val = $1.11x10^{-36}$) depending whether they overlap a designed sequence (control or ChIP peak) or not (random). Bar charts on the right in represents the distribution of significantly up-regulated regions from the ChIP peak category specifically, when considering an overlap with a ChIP peak of more or less than 60% (* = chi-square test p-val = 0.0003). **(b)** Distribution of STARR-seq regions according to the different categories, overlapping with the designed fragments (control or ChIP peaks) or randomly captured from the genome (random) **(c)** Bar chart showing the gradual decrease in recovered regions upon an increasing overlap with the designed region



Supplemental Figure S11 – STARR-seq distribution of expression. MAplot showing the distribution of STARR-seq fragment expression levels. Several subgroups can be distinguished: positive peaks (log2FC >1.5, p-val < 0.05, green), down (log2FC < -1.5, p-val < 0.05, red), negatives (log2FC < 0.0 and p-val > 0.1 or log2FC < 0.5 and p-val > 0.2, purple) or greyzone (not up or down-regulated or negative, yellow)



Supplemental Figure S12 – Behavior of control regions in STARR-seq. Promoters of housekeeping genes (HKG) (black triangle) and negative regions (red cross) both show no differential expression upon TP53 stimulation **(a, c).** Negative regions have a low basemean expression, while some HKG promoter have a higher basemean expression. This indicates the ability of the latter to drive expression independently of TP53 **(a-b).**

positives vs negatives

***de novo* length 19**

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)
1		1e-452	-1.042e+03	63.43%	0.46%	114.7bp (79.0bp)
2		1e-73	-1.681e+02	24.86%	1.82%	212.2bp (91.7bp)
3		1e-70	-1.618e+02	15.43%	0.50%	224.3bp (141.3bp)
4		1e-48	-1.122e+02	14.29%	0.90%	143.8bp (115.3bp)
5		1e-34	-7.965e+01	17.71%	2.59%	164.1bp (47.9bp)
6		1e-29	-6.870e+01	54.29%	25.64%	225.5bp (134.6bp)
7		1e-17	-3.937e+01	46.29%	25.02%	227.5bp (179.9bp)

p-value:	1e-452
log p-value:	-1.042e+03
Information Content per bp:	1.472
Number of Target Sequences with motif	222.0
Percentage of Target Sequences with motif	63.43%
Number of Background Sequences with motif	1.4
Percentage of Background Sequences with motif	0.46%
Average Position of motif in Targets	246.5 +/- 114.7bp
Average Position of motif in Background	159.4 +/- 79.0bp
Strand Bias (log2 ratio + to - strand density)	0.0
Multiplicity (# of sites on avg that occur together)	1.09
Motif File:	file (matrix) reverse opposite
PDF Format Logos:	forward logo reverse opposite

***de novo* length 20**

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)
1		1e-409	-9.436e+02	58.86%	0.27%	117.8bp (1.4bp)
2		1e-66	-1.538e+02	14.86%	0.35%	198.5bp (23.1bp)
3		1e-65	-1.499e+02	14.57%	0.49%	124.7bp (180.0bp)
4		1e-59	-1.377e+02	18.57%	1.27%	214.5bp (112.6bp)
5		1e-44	-1.032e+02	16.86%	1.37%	145.0bp (0.0bp)
6		1e-40	-9.340e+01	10.29%	0.50%	159.5bp (78.0bp)
7		1e-39	-9.135e+01	74.00%	39.13%	207.7bp (134.8bp)

p-value:	1e-409
log p-value:	-9.436e+02
Information Content per bp:	1.595
Number of Target Sequences with motif	206.0
Percentage of Target Sequences with motif	58.86%
Number of Background Sequences with motif	0.8
Percentage of Background Sequences with motif	0.27%
Average Position of motif in Targets	254.9 +/- 117.8bp
Average Position of motif in Background	64.5 +/- 1.4bp
Strand Bias (log2 ratio + to - strand density)	0.1
Multiplicity (# of sites on avg that occur together)	1.03
Motif File:	file (matrix) reverse opposite
PDF Format Logos:	forward logo reverse opposite

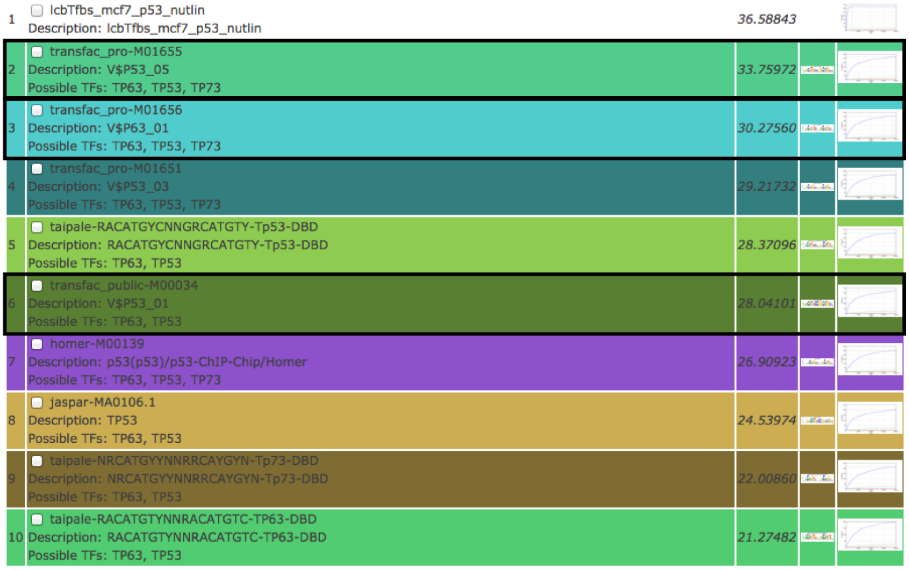
known motif

Total Target Sequences = 350, Total Background Sequences = 304

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif
1		p53(p53)/Saos-p53-ChIP-Seq/Homer	1e-322	-7.426e+02	0.0000	314.0	89.71%	20.1	6.62%
2		p63(p53)/Keratinocyte-p63-ChIP-Seq/Homer	1e-279	-6.431e+02	0.0000	310.0	88.57%	26.8	8.82%
3		p53(p53)/p53-ChIP-Chip/Homer	1e-272	-6.278e+02	0.0000	265.0	75.71%	14.3	4.71%
4		p53(p53)/mES-cMyc-ChIP-Seq/Homer	1e-106	-2.444e+02	0.0000	128.0	36.57%	8.7	2.86%
5		Tcfcp2l1(CP2)/mES-Tcfcp2l1-ChIP-Seq/Homer	1e-39	-8.986e+01	0.0000	35.0	10.00%	1.6	0.52%
6		Smad4(MAD)/ESC-SMAD4-ChIP-Seq(GSE29422)/Homer	1e-34	-7.842e+01	0.0000	161.0	46.00%	53.6	17.63%
7		Smad3(MAD)/NPC-Smad3-ChIP-Seq(GSE36673)/Homer	1e-32	-7.560e+01	0.0000	236.0	67.43%	108.4	35.68%

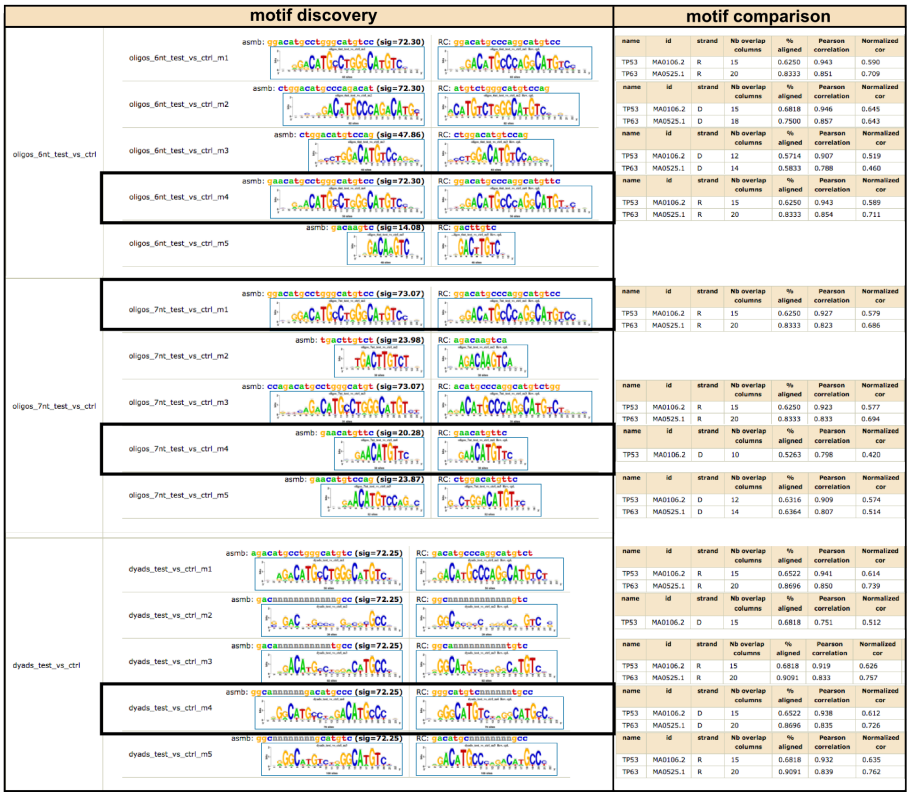
Supplemental Figure S13 – Results of HOMER motif discovery

positives vs negatives

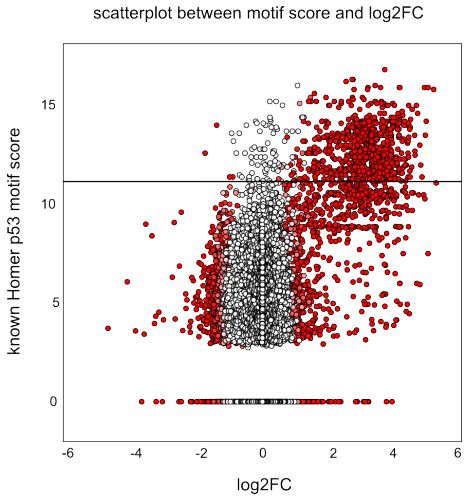


Supplemental Figure S14 – i-cisTarget results

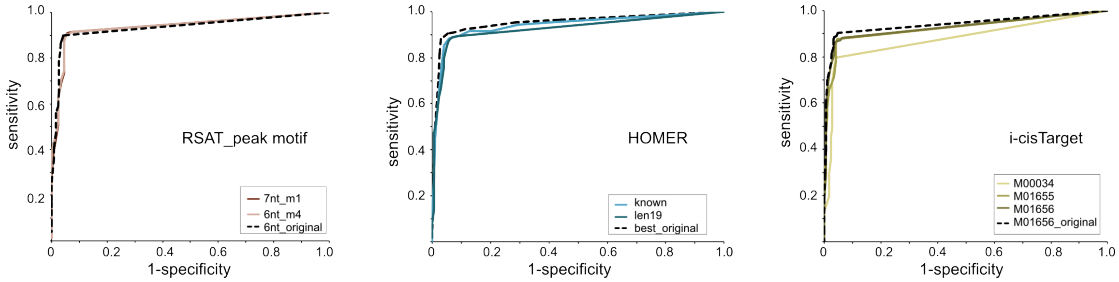
positives vs negatives



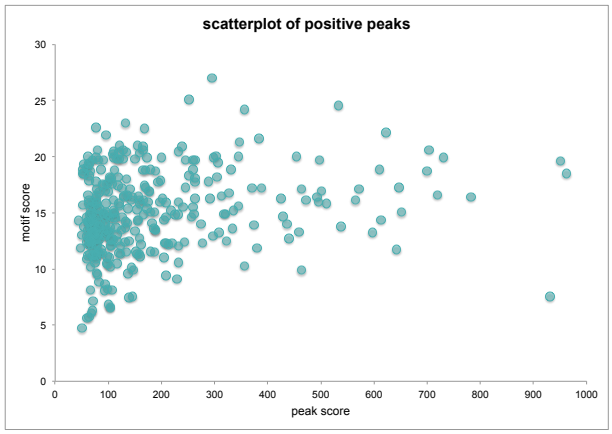
Supplemental Figure S15 – RSAT peak motif results



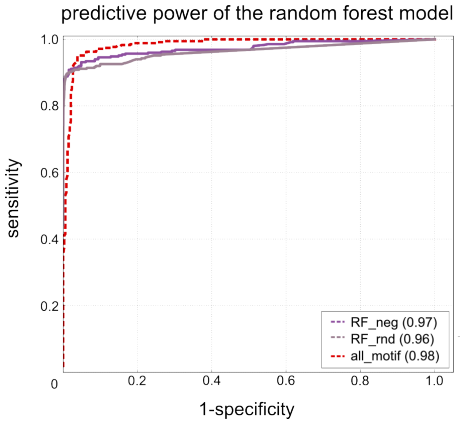
Supplemental Figure S16 – Scatterplot for motif score and log2FC of all CHEQ-seq regions. Each region is scored for the HOMER known TP53 motif and plotted against their barcode expression (log2FC). The minimal motif score for predictive power at specificity of 98% is indicated by a black horizontal line at score 11. Red color indicates significance (p-val < 0.05).



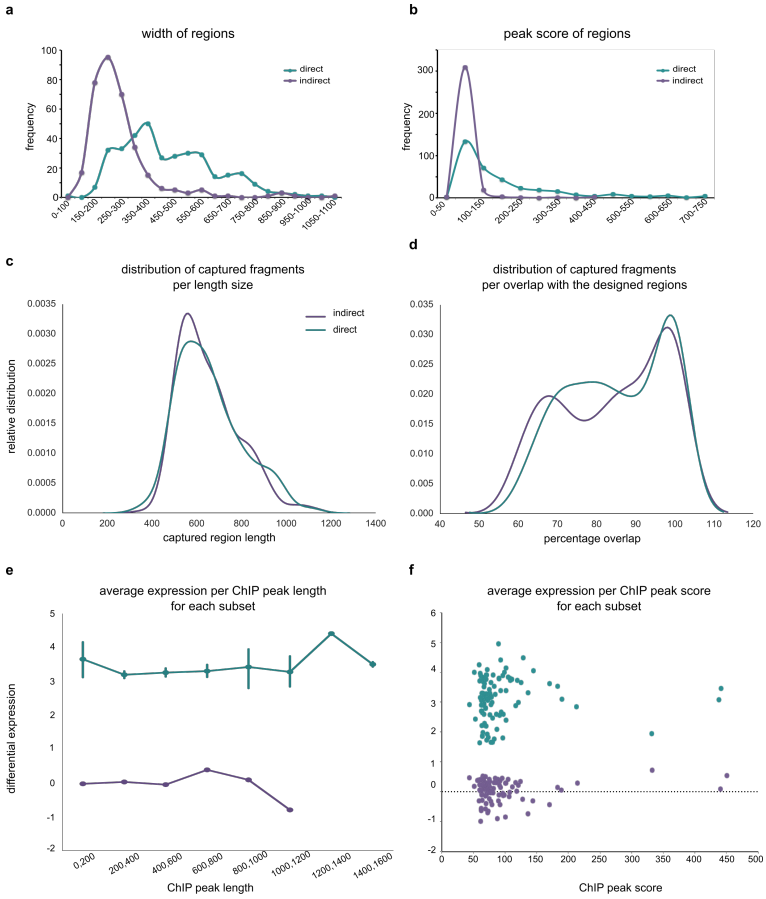
Supplemental Figure S17 – 2-fold crossvalidation of the motif recovery. Both the sets of positive and negative sequences were divided randomly in a training and test set. The following TP53 motifs were selected from motif discovery on the training set using the tools i-cisTarget, HOMER and RSAT peak motif. HOMER: a *de novo* motif length 19 (len19) and known motif from the HOMER collection ('known' (Koeppel et al. 2011)). i-cisTarget: three known Transfac motifs M01655, M01656 and M00034. RSAT peak motifs: 6 nucleotide (6nt_m4) and 7 nucleotide (7nt_1). The peaks in the test set were scored using Cluster-Buster (c and m = 0) for each individual motif. The best score for each peak was used to calculate the sensitivity and specificity for each motif. ROC curves are shown for RSAT-derived motifs, HOMER-derived motifs and i-cisTarget-derived motifs. A representative motif of the original recovery analysis, performed on the entire set of peaks is also included for each group (*_original).



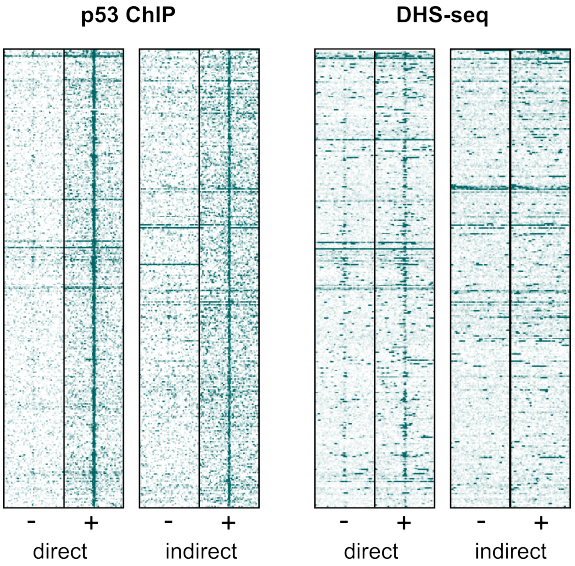
Supplemental Figure S18 – Comparison between motif score and peak score for the direct (positive) peaks.



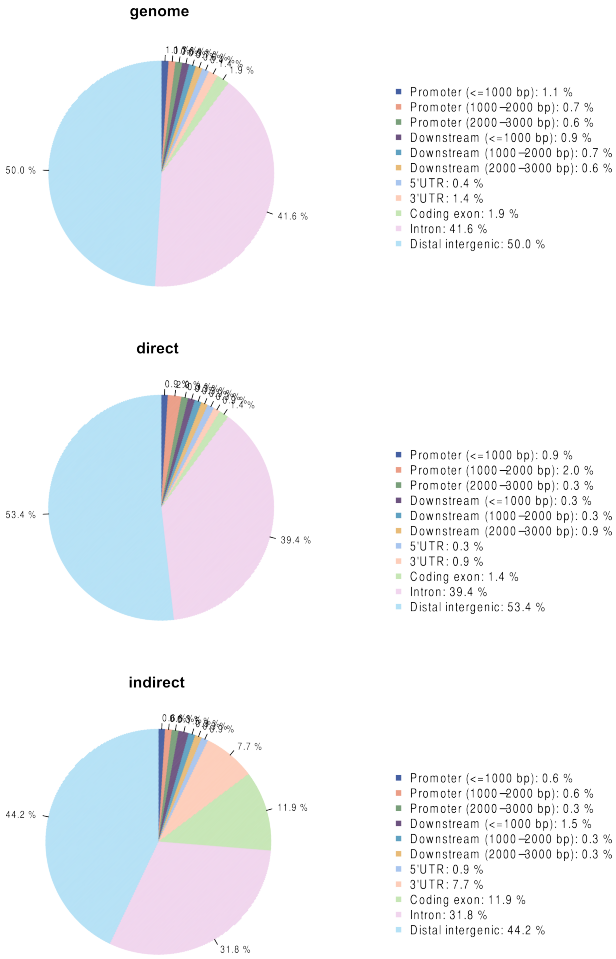
Supplemental Figure S19 – Random Forest model predictions. A Random forest model was trained using the CHEQ-seq positive regions. As a control random regions were selected from the genome (RF_rnd, purple-grey) or the CHEQ-seq negatives were used (RF_neg (bright purple), data cleaned for false negative predictions). The AUC for both models (AUC = 0.97 and 0.96) is comparable to that of the scoring with the combination of all motifs (AUC = 0.98) with a slightly higher specificity.



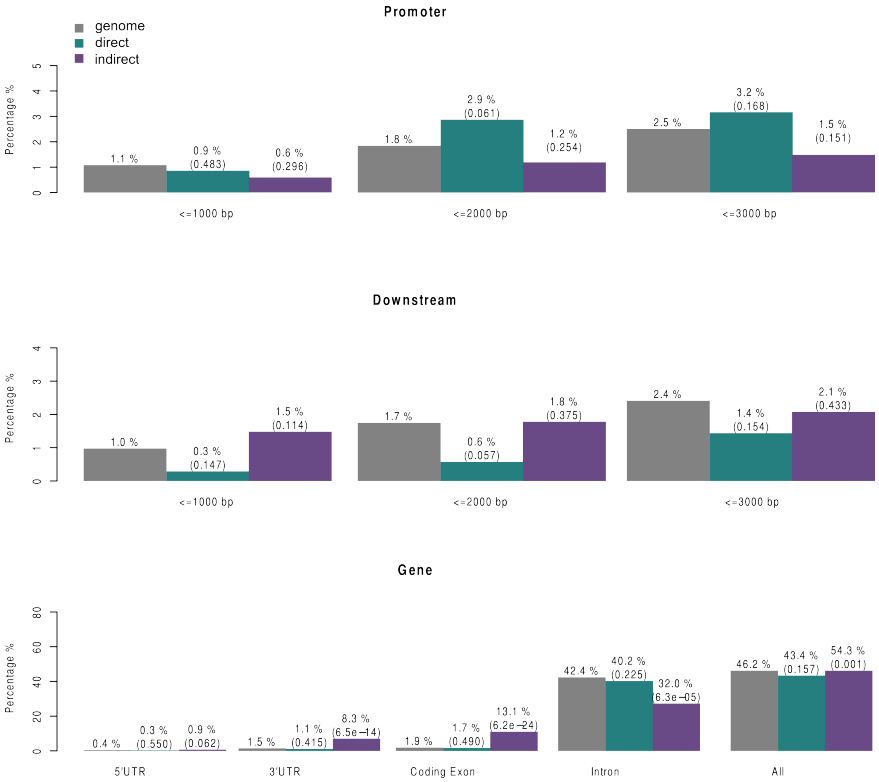
Supplemental Figure S20 – Comparison of peak length and score between direct and indirect peaks. (a) Distribution of peak width. Indirectly bound peaks (purple) are significantly shorter in peak width than directly bound peaks (green) (p-val = 6.06×10^{-42}). **(b)** Distribution of the peak score amongst peaks. Indirectly bound peaks (purple) have a significantly lower score than directly bound peaks (green) (p-val = 7.72×10^{-19}). **(c)** While the average peak width is different between the two sets, the size of the captured fragments is equally distributed between direct and indirect peaks. **(d)** The average distribution of the captured fragment width between the direct and indirect set does not differ, even when varying the percentage overlap of the captured regions with the design. **(e, f)** When subsampling direct or indirect peaks matching in width **(e)** or score **(f),** the expression patterns do not alter across the distribution of length or score within each set.



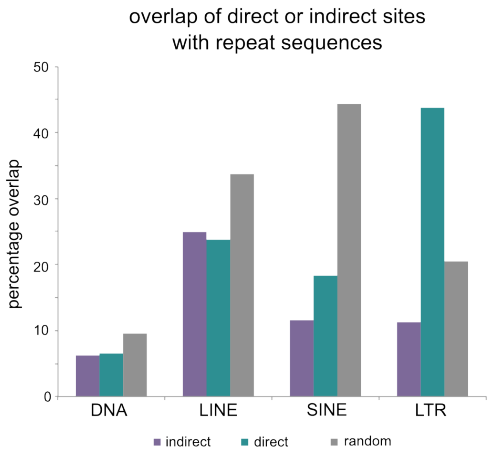
Supplemental Figure S21 – DHS-seq at direct and indirect peaks. Aggregation plot showing the DHS status of direct and indirect peaks. Upon Nutlin-3a stimulation (+), both direct and indirect peaks are bound by TP53 as determined by ChIP (left two plots) but only direct peaks become more open (right two plots). Peaks are extended to 2000bp each side.



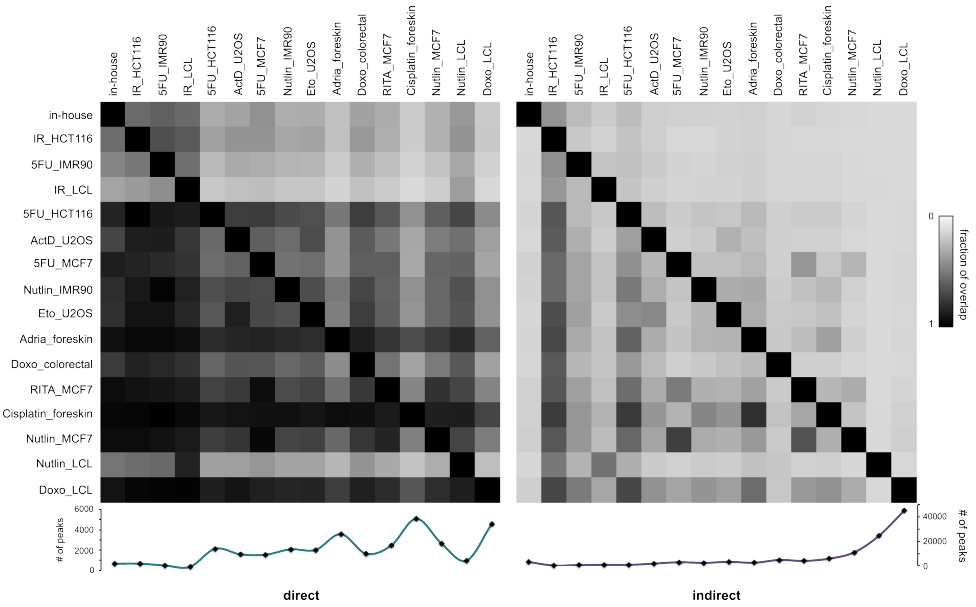
Supplemental Figure S22 – CEAS pie Charts for the distribution of regulatory regions. The distribution for various genomic elements showed a different distribution amongst the indirect peaks but not the direct peaks compared to the reference genome.



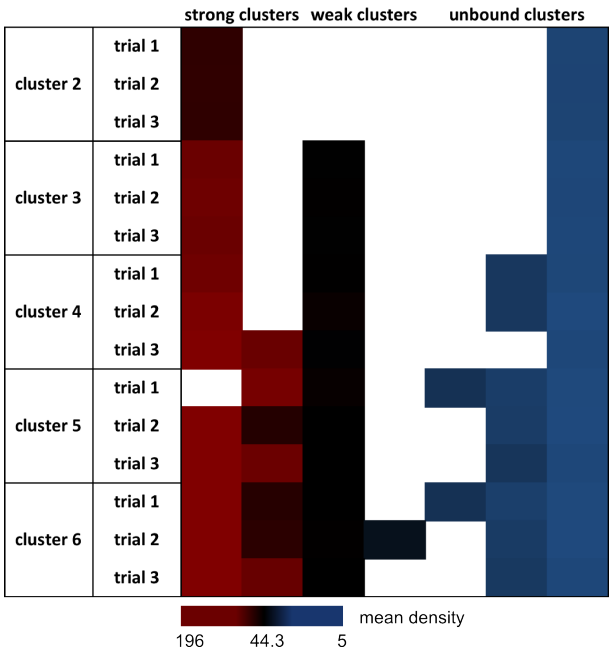
Supplemental Figure S23 – Distribution of CHEQ-seq peaks across the genome. Indirect peaks are signigicantly enriched in 3'UTR (p-val = 6.5×10^{-14}) and exonic (p-val = 6.2×10^{-24} regions compared to the reference genome.



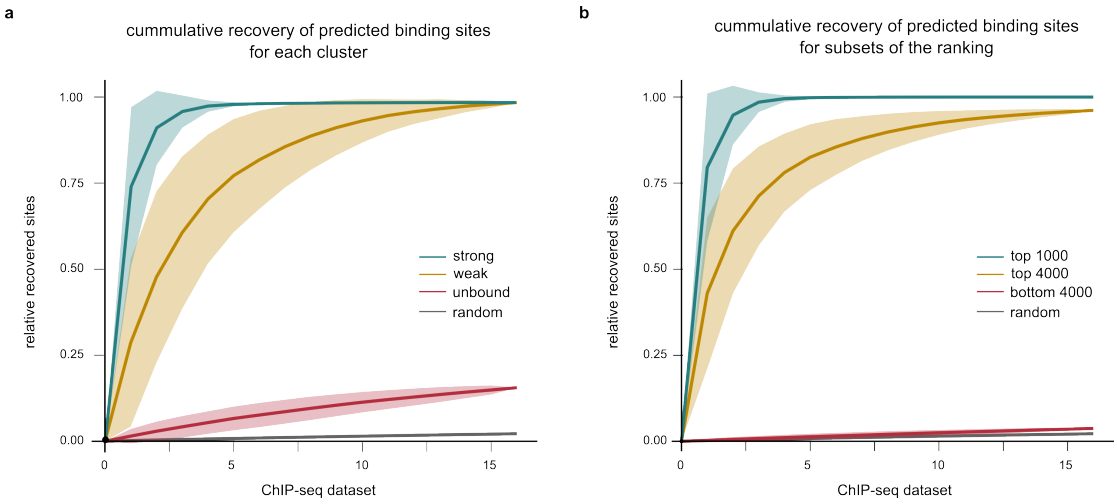
Supplemental Figure S24 – overlap of direct and indirect peaks with repeat sequences. Four different families of repeat sequences were collected from RepeatMasker: DNA transposons (DNA), long interspersed nuclear elements (LINE), Short interspersed nuclear elements (SINE) and long terminal repeats (LTR). 7000 random regions were selected in the genome as control. Significant Binomial values for indirect and direct: 0.00053 and 5.63E-05 (LINE), 2.2E-16 and 2.2E-16 (SINE), 1.02E-0.5 and 2.2E-16 (LTRs).



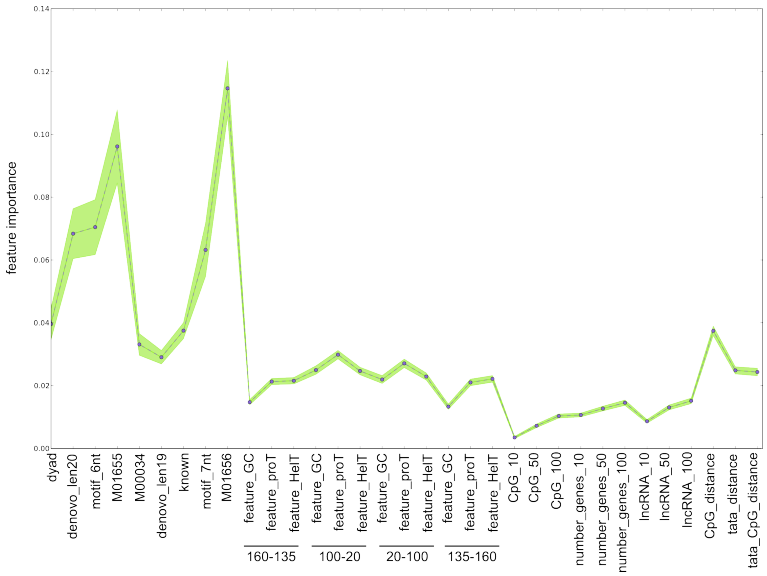
Supplemental Figure S25 – Comparison of direct and indirect peaks across all 16 ChIP-seq datasets. All 15 public datasets and the in-house dataset were compared to each other for their overlap in direct or indirect peaks. Each column represents the fraction of peaks of a particular dataset overlapping with each other set (row). Below each matrix the total number of direct or indirect peaks for each dataset is depicted.



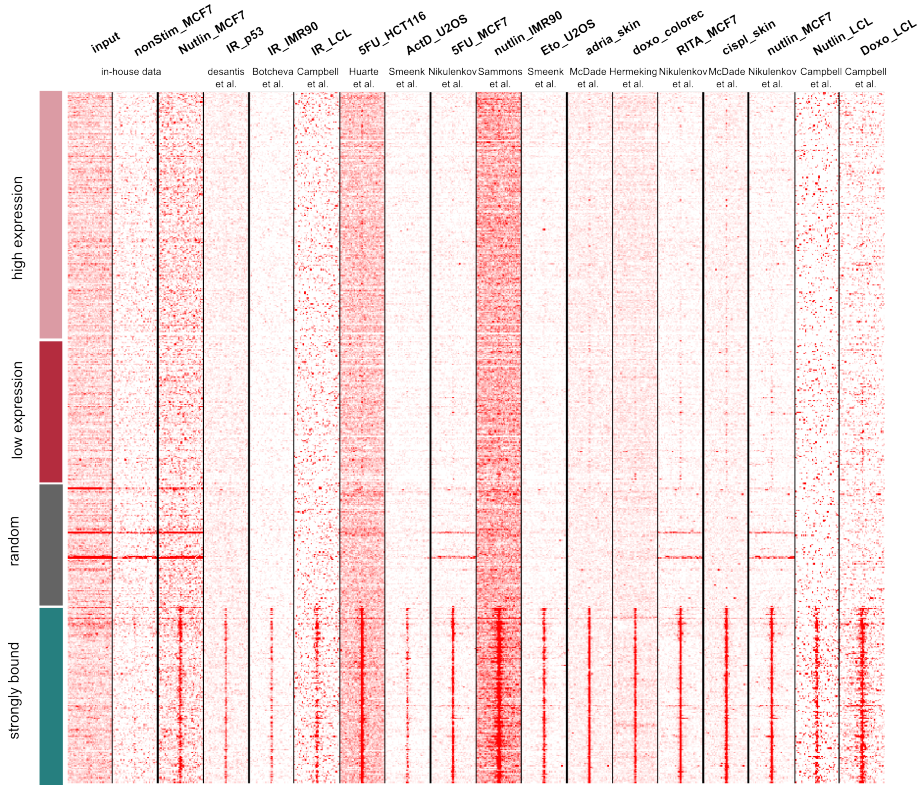
Supplemental Figure S26 – Analysis of different clustering parameters. Different settings for the k-means clustering were tested in SeqMiner, varying the k-value between 2 and 6. Each value was tested on three separate random seeds. The maximum mean density value for each obtained cluster was noted. Analyses with k > 3 generated several subclusters with very similar mean density values that mimicked the initial three clusters (i.e. strong, weak and unbound).



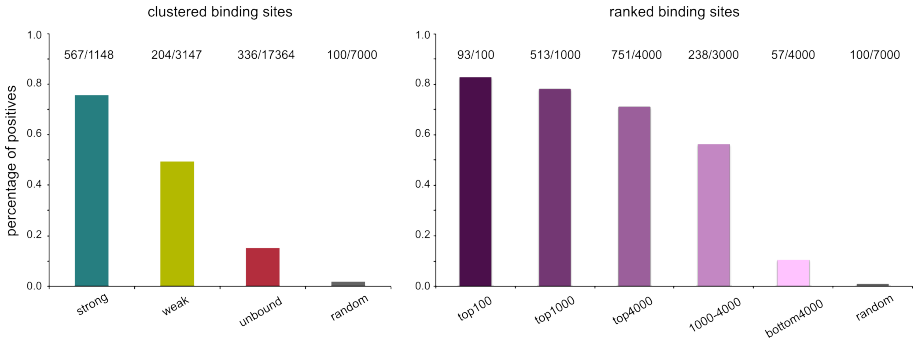
Supplemental Figure S27 – Recovery of predicted TP53 binding sites from all ChIP-seq datasets. The cumulative recovery of the predicted TP53 binding sites over the 16 datasets was checked either for the three clusters (left) or several sections of the ranked list (right). The strongly bound sites or top 1000 sites (green) could all be recovered using only 5 datasets. In contrast, the unbound sites or bottom 4000 sites (red) were hardly recovered with all datasets combined, with the bottom 4000 show little difference to a set of 7000 random genomic regions (grey). The weak sites or the top 4000 sites (yellow) show a more modest recovery, where all sites are retrieved, but requiring all datasets. The order of the datasets was shuffled randomly over 1000 iterations.



Supplemental Figure S28 – Feature importance for the random forest model. The contribution of each feature to the predictive power of the Random Forest model was plotted, showing that most contribution comes from the motifs, in particular the two Transfac motifs M01655 and M01656. The features are as follows: (1-9) represent the selected TP53 motifs. (10-21) the DNA shape features at different flanking regions of the binding site are plotted (from 5' to 3'). (22-30) the number of CpG islands, genes or lncRNA in different regions around the sites is presented (10kb, 50kb and 100kb). (31-33) The distance to CpG or tata promoters or both is plotted.



Supplemental Figure S29 – Coverage signal amongst subsets of unbound sites. Unbound sites represented by a CHEQ-seq captured region were distributed based on their reporter expression level into low expressed (216 sites) and high expressed (120 sites, dark red). As comparison, 100 randrom regions form the genome (grey) and 150 randomly selected strongly bound sites (green) were also plotted. Neither high nor low expressing unbound sites show any TP53 binding indicating they were all correctly assigned as unbound sites despite their differences in reporter activity.



Supplemental Figure S30 – Testing of predicted TP53 binding sites with MPRAs. TP53 binding sites clustered (left) or ranked (right) that were captured and tested in either CHEQ-seq or STARR-seq were analysed for differential expression. Sites in a captured region with log2 fold change > 1.5 and p-val <0.05 were termed as positive.