CRISPR/Cas9 screen for functional MYC binding sites reveals MYC-dependent vulnerabilities in K562 cells

Marta Kazimierska[1], Marta Podralska[1], Magdalena Żurawek[1], Tomasz Woźniak[1], Marta Elżbieta Kasprzyk[1], Weronika Sura[1], Wojciech Łosiewski[1], Iwona Ziółkowska-Suchanek[1], Joost Kluiver[2], Anke van den Berg[2], Natalia Rozwadowska[1], Agnieszka Dzikiewicz-Krawczyk[1]*

[1]Institute of Human Genetics, Polish Academy of Sciences, Poznań, Poland

[2]Department of Pathology and Medical Biology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

*correspondence: agnieszka.dzikiewicz-krawczyk@igcz.poznan.pl; Strzeszyńska 32, 60-479 Poznań, Poland

ABSTRACT

The transcription factor MYC is a proto-oncogene with a well-documented essential role in the pathogenesis and maintenance of several types of cancer. MYC binds to specific E-box sequences in the genome to regulate expression of adjacent genes in a cell type- and developmental stage-specific manner. To date, a comprehensive analysis of direct MYC targets with essential roles in different types of cancer is missing. To enable identification of functional MYC binding sites and corresponding target genes, we designed a CRISPR/Cas9 library to destroy E-box sequences in a genome-wide fashion. As a proof of principle, using this library we identified several known and novel MYC targets critical for K562 chronic myelogenous leukemia cells and uncovered specific features of essential E-boxes. Our unique, well-validated tool opens new possibilities to gain novel insights into MYC-dependent vulnerabilities in any cancer type.

INTRODUCTION

MYC is an oncogene broadly involved in many tumors. Due to amplifications, mutations, translocations or posttranslational modifications, MYC is highly expressed in up to 70% of cancers[1]. The family of MYC transcription factors consists of three members: C-MYC, N-MYC, and L-MYC[2]. Among them, C-MYC (further called MYC) has the strongest oncogenic potential and is widely deregulated in cancer, while N-MYC and L-MYC are mainly involved in neuroblastoma and lung cancer, respectively[3–5]. MYC regulates expression of genes associated with cell cycle, apoptosis, proliferation, cellular differentiation as well as strongly alters metabolism of cancer cells and stimulates ribosome and mitochondrial biogenesis[6–9]. Altogether, this creates strong addiction of cancer cells to MYC and it has been demonstrated that indeed MYC withdrawal leads to tumor regression[10]. As such, targeting MYC

appears as an attractive strategy for cancer therapies. However, no clinically relevant MYC-targeting therapies have been developed so far[11]. MYC is commonly thought to be an "undruggable target" due to its localization and activity in the nucleus and lack of an active site for interaction with small molecules[2,4]. Therefore, there is a need to look for indirect approaches such as identification of essential MYC-regulated genes which may serve as a basis for novel anti-tumor therapies.

MYC is a transcription factor (basic helix-loop-helix-leucine-zipper, bHLH-LZ) that creates heterodimers with the MYC-associated factor X (MAX). The MYC/MAX complex can recognize and bind to so-called E-box motifs in DNA (canonical sequence 5'-CACGTG-3'), localized mainly in promoters and enhancers. MYC activates transcription by releasing the paused RNA polymerase II and by recruiting chromatin-modifying complexes such as TIP48, TIP60, TRRAP or GCN5[2]. On the other hand, by competing with other transcription factors such as MIZ1 or SP1, MYC can also repress transcription[12]. Although MYC has been widely studied, its targetome is not fully known. Thousands of genes responding to MYC activation or inhibition have been identified but it is difficult to distinguish direct and indirect targets[13–16]. Chromatin immunoprecipitation with MYC antibody followed by sequencing (MYC-ChIP-seq) indicated thousands of MYC-bound sites in the genome[13,17–22]. However, the set of MYC target genes varies among different cell types and developmental stages[23,24]. This may be explained by the fact that MYC binds predominantly to already active promoters or enhancers and inactive genes remain silent[17,22]. Altogether, it is still not clear which MYC targets are essential for cancer cells. To date, a limited number of RNA interference screens for genes essential in MYC-driven tumors have been performed. However, they either focused only on a limited set of genes or were performed in normal cells with forced MYC overexpression[25–29]. This may preclude their direct relevance for cancer cells. So far, no global, comprehensive analysis of MYC target genes essential for MYC-dependent cancer cells has been performed.

In this study, we set forward to create a tool for a comprehensive analysis of essential MYC binding sites and target genes in MYC-dependent cancer cells. We designed a sgRNA library for a genome-wide disruption of MYC binding sites. As a proof of principle, we conducted a high-throughput screen with the MYC-CRISPR library in the MYC-dependent chronic myelogenous leukemia (CML) cell line K562[30,31]. In parallel, we utilized the Brunello library for genome-wide knockout of protein-coding genes. Overlapping data from both screens identified several known and novel MYC targets critical for K562 cells, and indicated specific features of essential E-boxes. Altogether, we established a unique, well-validated tool to identify MYC-regulated target genes relevant for growth of malignant cells. Our findings provide novel insights into MYC-dependent vulnerabilities in K562 cancer cells.

RESULTS

Generation of the MYC-CRISPR library for a genome-wide disruption of MYC binding sites

From publicly available MYC-ChIP-seq data in MYC-dependent K562, MCF7, HepG2[32], and Burkitt lymphoma (BL) cell lines[20] we obtained 58,503 MYC binding sites, which contained 43,153 E-box motifs. 2,208 of them were located in the coding exons and were excluded to prevent identification of effects due to disruption of the protein. Based on the presence of PAM sequence, we were able to design 56,688 unique sgRNAs targeting the remaining 26,653 E-boxes (65.1%). After excluding sgRNAs with predicted off-target binding, we finally obtained 45,350 sgRNAs targeting 24,981 E-boxes (61%) (Figure 1A, Supplementary Table 1). Half of the E-boxes were targeted by more than one sgRNA (Figure 1B). The majority of E-boxes were located in introns (58.4%) or intergenic regions (33.2%). A small proportion of the E-boxes were present in 5' and 3' UTRs and non-coding genes (8.2%) or at exon-intron boundaries (0.2%) (Figure 1C). 61% of the E-boxes targeted by our MYC-CRISPR library were bound by MYC only in one cell line, while 6% of the E-boxes were commonly bound by MYC in all four cell lines (Figure 1D). This is in line with the high cell-type specificity of MYC targets. Although the MYC-CRISPR library was designed based on the data from selected four cell lines, we also checked its compatibility with other cell lines. Analysis of available MYC-ChIP-seq data from 11 cell lines revealed that 16-60% MYC binding sites were targeted by the library sgRNAs (Figure 1E). This indicates that the MYC-CRISPR library can be widely applied for studies in various cell lines.

Next generation sequencing of the plasmids obtained after cloning revealed high quality of the MYC-CRISPR library. All sgRNAs were present in the library and the skew ratio of 90th to 10th percentile was only 1.76, indicating a uniform representation of all constructs in the library (Figure 1F). The quality of the amplified Brunello library also conformed to the recommended requirements (Supplementary Figure 1A). Thus, we generated a high quality MYC-CRISPR library for genome-wide targeting of MYC binding sites that can be universally used in various cell lines.

High-throughput screen for functional MYC binding sites and target genes essential for growth of K562 cells

To identify essential MYC binding sites and target genes, we conducted a genome-wide high-throughput CRISPR/Cas9 knockout screen with the MYC-CRISPR library (46,354 sgRNA) targeting E-boxes and the Brunello library (77,441 sgRNA)[33] targeting protein-coding genes (Figure 2A). K562 cells were infected in duplicate aiming at ~500x coverage of each sgRNA in both libraries and a 30% transduction efficiency. The actual coverage for the MYC-CRISPR library was 455-540x with transduction efficiency ranging from 26.9-32%, while 23-24.5 % of cells were transduced with the Brunello library, resulting in 390-415x coverage (Figure 2B). Cells were collected at T0 (after puromycin

selection) and T1 (after 20 population doublings) and next-generation sequencing was performed to determine the abundance of sgRNA constructs (Figure 2C, Supplementary Table 2, 3). None of the non-targeting sgRNAs was depleted ≥2-fold in both replicates and 2-4 sgRNAs targeting MYC showed consistently decreased abundance at T1. 11,106 sgRNAs from the Brunello library and 1,663 sgRNAs from the MYC-CRISPR library were consistently ≥2-fold depleted in both replicates (Figure 2D). Considering the combined effect of all sgRNAs targeting a given gene or E-box, using DESeq2 algorithm, 1,226 genes (Supplementary Table 4) and 152 E-boxes (Supplementary Table 5) were identified as essential for growth of K562 cells (p_{adj}<0.001), while 5 E-boxes and 10 genes were significantly enriched (Figure 2E). Analysis of the predicted sgRNA cutting efficiency by WU-CRISPR bioinformatic tool[34] revealed that both in the MYC-CRISPR and Brunello libraries essential sgRNAs had significantly higher scores than non-essential ones (Supplementary Figure 1B, 1C). Thus, we cannot exclude that some E-boxes and genes were not identified as essential in our screens due to lower efficiency of DNA editing by sgRNAs.

We noticed that 40% of the essential E-boxes were localized on chromosome 22, near the breakpoint in *BCR* involved in the t(9;22) translocation leading to the *BCR-ABL1* fusion (Figure 2F). Hits observed in this region are most likely not caused by targeting essential genes but due to the presence of a tandem amplification of this region in K562 cells, as Wang et al. observed previously[35]. Massive CRISPR/Cas9-mediated DNA cleavage is toxic for the cells regardless of the gene/regulatory region targeted. Indeed, an orthogonal approach with dCas9 which does not induce DNA cleavage but blocks E-box sites to prevent MYC binding demonstrated no effect on K562 cell growth for two sgRNAs from chromosome 22q11 (Supplementary Figure 2A). Therefore E-boxes from the amplified region on 22q11 and 9q34 were excluded from further analysis. For 19% of the remaining 89 E-boxes, adjacent genes were also essential in the genome-wide Brunello screen, while 42% of these genes were well-known MYC targets (Supplementary Table 5). The essential E-boxes tended to be close to essential genes, and essential genes were enriched near essential E-boxes (Figure 2G). Thus, our CRISPR/Cas9 screen revealed known and novel MYC-dependent vulnerabilities in K562 cells.

Validation of the CRISPR/Cas9 screen results

To validate the results of the screen and confirm robustness of our approach for identification of MYC-dependent vulnerabilities in cancer cells, we focused on the top 10 most significantly depleted E-boxes. Of these, we included for validation six that were located nearby a protein-coding gene that was at least 4-fold depleted in Brunello screen (chr3_BS897_CATGTG, chr11_BS79_CGCGTG, chr11_BS2113_CACATG, chr13_BS121_CGCGTG, chr17_BS377_CACGTG, chr19_BS2255_CACATG), and two E-boxes adjacent to long non-coding RNA genes (chr10_BS212_CACATG,

chr2_BS1664_CACGTG) (Figure 2F). The remaining two of top 10 E-boxes were in vicinity of non-essential genes and were not considered for validation. In addition, we included a non-essential E-box, chr17_BS377_CGCGTG which was located 12 nt downstream of chr17_BS377_CACGTG (Supplementary Figure 3B), for validation as well to gain further insight into MYC regulation at this locus.

Cells were transduced with individual sgRNAs targeting selected E-boxes. First, we checked whether our approach allows for efficient disruption of E-box motifs. TIDE analysis confirmed DNA editing with >90% efficiency for all sgRNAs. The spectrum of mutations varied between individual constructs, with small 1-2 nt insertions and deletions being most prevalent (Figure 3A).

Next, we aimed to establish whether disruption of E-boxes affected expression of adjacent genes. Results of qRT-PCR showed that for 6 out of 8 selected E-boxes expression of at least one gene was significantly decreased (Figure 3B). In some instances, expression of both genes nearby an E-box was affected, while in others one of the genes was not affected at all or to a lesser extent. Interestingly, for chr17_BS377 we observed strong downregulation of *PFAS* when targeting the CACGTG E-box (sg1), but much weaker effect when targeting the CGCGTG E-box (sg2 and sg3). This is in line with the fact that we only identified CACGTG as essential in our screen. For E-box chr2_BS1664 we could not reliably detect expression of the adjacent lncRNAs, neither in control nor CRISPR/Cas9 edited samples. For chr11_BS2113 the closest gene was *PRKRIR* located >50 kb, and we did not observe an impact on its expression. Thus, disruption of E-boxes with CRISPR/Cas9 modulates expression of target genes and can indicate genes regulated by MYC within a given locus.

We further validated the effect on cell growth observed in high-throughput screens for three E-boxes with the strongest effect on expression of adjacent genes: chr10_BS212, chr11_BS79 and chr17_BS377. We conducted growth assays in K562 cells transduced with sgRNAs for E-box disruption and knockout of adjacent genes, and shRNAs for knockdown of the long non-coding RNA PRKCQ-AS1, which was not included in the Brunello screen. For all sgRNAs targeting chosen E-boxes we observed a significant decrease in cell growth, consistent with the results of the screen with the MYC-CRISPR library. We confirmed that within chr17_BS377, only the CACGTG E-box targeted by sg1 is essential for K562 cells growth. Moreover, knockout/knockdown of genes adjacent to each E-box also significantly reduced cell growth, in line with the effect observed in the Brunello screen (Figure 3C). Within chr10_BS212 both *PRKCQ* and *PRKCQ-AS1*, and within chr11_BS79 both *PIDD1* and *RPLP2* decreased expression after E-box disruption, confirming that they are regulated by MYC. The individual knockout or knockdown of these MYC regulated genes allowed pinpointing the MYC targets relevant for the cell growth.

To confirm direct MYC binding and regulation of transcription, we performed a luciferase reporter assay for the three selected E-boxes. We observed decreased luminescent signal for all three MYC binding sites upon MYC knockdown (KD) (Figure 3D). This indicates that binding of MYC to these sequences results in MYC-dependent transcription. Moreover, MYC-ChIP in cells transduced with sgRNAs targeting two of the selected E-boxes confirmed decreased MYC binding to chr17_BS377 and chr10_BS212 as compared to non-targeting control sgRNA. Disruption of chr11_BS79 did not affect the strength of MYC binding (Figure 3E).

As an alternative approach, to further validate the importance of MYC binding we utilized dCas9 to block the E-box sequences rather than to disrupt them when using WT Cas9. RT-qPCR in cells transduced with dCas9 and three sgRNAs targeting chr10_BS212, chr11_BS79 and chr17_BS377 showed the same pattern of gene expression as for WT Cas9 (Supplementary Figure 2B). In addition, the effect on cell growth for chr11_BS79 and chr17_BS377 was similar with dCas9 and WT Cas9 (Supplementary Figure 2A) while we did not notice differences in cancer cell growth compared to the non-targeting control for the sgRNA targeting chr10. These results further confirm that disturbing MYC binding at these positions is causative for the observed effects on expression of target genes and cell growth.

Altogether, we confirmed that our approach allows for efficient disruption of E-box motifs, which results in decreased MYC binding and affects expression of target genes.

Identification of critical hubs in MYC-regulated pathways

Gene ontology (GO) analysis of essential genes from the Brunello screen revealed enrichment of processes crucial for cancer cells such as ribosome biogenesis, RNA splicing, translation, cell cycle, RNA biosynthesis, and others (Figure 4A). To gain insight into the processes regulated by MYC, we focused on the essential genes within 50 kb from the E-boxes which were at least two-fold depleted in the screen. GO analysis indicated enrichment of processes known to be activated by MYC, such as ribosome biogenesis, nucleic acid metabolism, translation, and metabolism (Figure 4B). Similar results were obtained with Gene Set Enrichment Analysis (GSEA). Interestingly, for both libraries the 'HALLMARK_MYC_TARGETS' gene set was significantly enriched (Supplementary Figure 4), which confirms that our approach can accurately identify essential MYC-dependent genes.

One of the top essential E-boxes, chr17_BS377, was localized in the 5'UTR of *PFAS* gene encoding phosphoribosylformylglycinamidine synthase. This enzyme is involved in the purine biosynthesis pathway which generates inosine 5'-monophosphate (IMP) from phosphoribosyl pyrophosphate (PRPP) and is crucial for cancer cell proliferation and viability[36]. Six enzymes catalyze the ten steps of PRPP to IMP conversion: PPAT, GART, PFAS, PAICS, ADSL and ATIC (Figure 4C), all of them being known

MYC targets. Results of our Brunello screen indicated that five enzymes (except ADSL) were significantly or at least four-fold depleted, confirming their essentiality in K562 cells. All genes had at least one E-box nearby targeted by the MYC-CRISPR library. However, only disrupting the E-box in *PFAS* inhibited K562 cell growth in our screen. Hence, we further investigated the role of MYC in the purine biosynthesis pathway. First, we confirmed with TIDE analysis that the E-boxes near *PPAT, GART, PFAS, PAICS,* and *ATIC* were efficiently disrupted (78-100%) (Figure 4D). Then we looked at the expression of those five genes upon E-box disruption. *GART* levels were decreased up to ~60% by both sgRNAs targeting E-boxes within the gene. *PPAT* and *PAICS* are located next to each other in a head-to-head orientation, with three E-boxes in the 5'UTR and intron of *PAICS*. Neither of the E-boxes affected the expression of *PPAT*, while disruption of two intronic E-boxes reduced *PAICS* levels by up to 30%. Similarly, targeting two E-boxes in the 5'UTR of *ATIC* decreased its expression by 40%. However, the strongest downregulation was observed for *PFAS* with >90% reduced expression (Figure 4D). Taking into account that of all these genes *PFAS* mRNA levels are the lowest, this suggests that PFAS may be the rate-limiting enzyme in the purine biosynthesis pathway and the critical MYC-regulated hub, in line with our observation that only the E-box in *PFAS* was essential for K562 cells.

Another example of a pathway identified in our study with multiple MYC-dependent elements is the spliceosomal U1 small nuclear ribonucleoprotein (snRNP) core formed by 7 Sm proteins: SNRPB, SNRPD1, SNRPD2, SNRPD3, SNRPE, SNRPF and SNRPG (Figure 4E). Apart from *SNRPB* and *SNRPE,* the other genes encoding these proteins were identified as essential in our Brunello screen. Moreover, all of them have been previously indicated as MYC targets and, except for *SNRPB* and *SNRPG,* there was an E-box nearby targeted by the MYC-CRISPR library. Still, only disrupting the E-box in *SNRPD2* had an adverse effect on K562 cells. This may indicate that SNRPD2 is the critical MYC-regulated player in splicing. The third example of a pathway with multiple MYC-regulated components is the mitochondrial fatty acid synthesis with five proteins: MECR, MCAT, OXSM, RPP14 and HSD17B8 (CBR4) that catalyze fatty acid chain elongation. Although only *MECR, RPP14* and *MCAT* are known MYC targets, all genes had an E-box nearby which was included in the MYC-CRISPR library. Moreover, all genes except for *HSD17B8* were essential for K562 cells. Again, however, only disrupting the E-box in the 5'UTR of *MECR* affected cell growth (Figure 4F), which suggests MECR as the crucial MYC-dependent link in the mitochondrial fatty acid synthesis.

Features of the essential E-boxes

To further characterize the 89 essential E-boxes we looked into several features, such as strength of MYC binding, presence of epigenetic marks and expression of adjacent genes, based on publicly available datasets.

MYC-ChIP signals for essential E-boxes were slightly stronger as compared to non-essential E-boxes (median value 241 vs 218 for non-essential) but the difference was not statistically significant (Figure 5A). Localization within DNase hypersensitivity regions, indicating open chromatin, did not discriminate essential E-boxes from non-essential (Figure 5B). We did not observe a correlation between H3K27ac and H3K4me1 (mark of active enhancers) in either essential or non-essential E-boxes. On the other hand, E-boxes were enriched in regions co-occupied by H3K27ac and H3K4me3 (mark of promoter regions), with a significantly stronger correlation between both histone marks for essential vs non-essential E-boxes (Pearson r^2=0.73 and r^2=0.44, respectively) (Figure 5C, 5D).

Expression of genes localized near essential E-boxes was significantly higher compared to genes near non-essential E-boxes (Figure 5E). Similarly, expression of essential genes in the Brunello library was significantly higher compared to non-essential genes (Figure 5E). The fact that essential genes are highly expressed is rather intuitive; nonetheless, the observation that essential E-boxes tend to be near highly expressed genes is useful for prediction of E-boxes relevant for cancer cells.

Altogether, we conclude that essential E-boxes tend to be localized in promoter regions of highly expressed genes, while strength of MYC binding and DNase hypersensitivity is not related to the relevance of E-box for K562 cells.

Implications of the sequence context on E-box functionality

We observed that in some cases with a +1 insertion, the inserted nucleotide did not change the E-box motif (i.e. G added before the last G in the E-box or C inserted after the first C in the E-box). For some sgRNAs the estimated frequency of such DNA edits reached up to ~50% (Supplementary Table 6). Despite not clearly affecting the E-box sequence we did observe an effect on cell growth and expression of adjacent genes on bulk infected K562 cells. To gain further insights into the E-box grammar, we focused on the E-box CACATG on chromosome 10 (chr10_BS212) with a cut site directly before the last G in the E-box (CACAT*G) and the highest percentage (56%) of +1 insertions with an additional G (CACGTGG).

We successfully established 24 clones, among which were cells with WT sequence on all 3 alleles (K562 cells are triploid), cells with a mixture of WT and various indel alleles, and three clones homozygous for the +G insertion on all 3 alleles (Supplementary Table 7). The RT-qPCR analysis did not reveal significant differences between WT homozygotes and mutants in the expression levels of the nearby PRKCQ gene, that was not essential for K562 cells in the Brunello screen. In contrast, expression of the other adjacent gene, long non-coding RNA PRKCQ-AS1, whose downregulation negatively affected K562 cell growth, was significantly decreased in +G homozygotes, to a similar extent as in clones with

other indels which clearly disrupted the E-box (Figure 6A). This indicated that even though the E-box motif sequence *per se* was not changed, its functionality was affected.

This prompted us to look at the sequences flanking the E-boxes. Analysis of the nucleotide frequency 20 nt upstream and downstream of non-essential E-boxes showed uniform distribution of nucleotides. In contrast, there were marked differences at particular positions flanking essential E-boxes (Figure 6B). Statistical analysis of 10 nt upstream/downstream revealed that certain nucleotides were significantly over- or underrepresented in the neighborhood of essential E-boxes (Figure 6C). Since E-boxes are (quasi)palindromic and can be read on either strand, G at +1 equals C at -1 etc. The strongest bias in nucleotide composition was observed at positions +/-4 nt, 5 nt and 7 nt. Moreover, we observed a strong overrepresentation of either a T immediately after or an A immediately before essential E-boxes and a trend for reduced frequency of G immediately after/C immediately before essential E-boxes. When we considered both immediate adjacent nucleotides before and after E-box simultaneously, the configuration A-Ebox-G or C-Ebox-T was two times more frequent among essential E-boxes, while C-Ebox-G occurred three times less among essential E-boxes (Figure 6D). Thus, we speculate that the immediate context of E-boxes might affect their functionality. This could explain why such changes without apparent disruption of the E-box affected expression of adjacent genes and cell viability.

Discussion

Despite decades of research, MYC still evades full comprehension and therapeutic targeting. To understand the mechanisms underlying cancer cell addiction to MYC, it is essential to determine the crucial genes regulated by this transcription factor. Therefore, the aim of this study was to identify on a genome-wide scale functional MYC binding sites and the corresponding target genes essential for K562 cells growth. To this end, we have established a novel CRISPR/Cas9-based tool to disrupt MYC-bound E-boxes. Using this new approach, we were able to determine MYC-dependent vulnerabilities in K562 cells, including multiple novel MYC targets.

A complete view of a transcription factor action requires the knowledge of their target genes together with binding partners and the mechanisms involved in the recognition of specific binding sites[37,38]. Several methods with various pros and cons enable studying TF function, e.g. gene expression analysis upon TF knockout or chromatin immunoprecipitation. Recently, pooled CRISPR/Cas9 screens have emerged as a powerful tool in molecular biology and have been also applied for TFs analysis. Different variants such as CRISPRa/CRISPRi, CRISPR/dCas9 allow for identification and analysis of target genes as well as regulatory elements such as enhancers and promoters[39–44]. So far, wild-type Cas9 and dCas9 have been mostly used for studying TF binding sites. WT Cas9 disrupts TF binding by introducing

random indels and if the recognition sequence is gone, the TF can no longer bind. However, in case of longer and degenerate motifs, small indels may not be enough to disrupt them. Moreover, this approach cannot be applied to motifs localized in coding exons or in highly amplified regions. On the other hand, dCas9 does not introduce double-strand breaks into the DNA but blocks the target motif and prevents binding of other proteins[45–47]. However, dCas9 should not be applied in promoter regions, as it may interfere with the whole transcription machinery rather than specifically block TF binding.

Extensive validation of our approach confirmed efficient E-box disruption, with predominance of +1 nt insertions and -1 nt deletions. Furthermore, we showed that E-box disruption significantly decreased expression of adjacent genes and that validated E-boxes were regulated by MYC. MYC-ChIP revealed decreased MYC binding for two out of three selected E-boxes upon CRISPR/Cas9 editing. The lack of effect for the E-box on chr11 can be explained by the presence of another, non-essential E-box, chr11_BS79_CACGCG ~100 bp downstream of the analyzed E-box chr11_BS79_CGCGTG (Supplementary Figure 3A). Resolution of ChIP-PCR with DNA fragments of ca. 500 bp does not allow to distinguish such close binding sites and this may explain why we did not observe reduced MYC binding upon disruption of the essential E-box chr11_BS79_CGCGTG. Finally, using individual sgRNAs we confirmed that E-box disruption or knockout of adjacent genes significantly decreased K562 cell proliferation, in line with the results of the screens. However, we observed some discrepancies using a parallel dCas9 approach to target E-boxes. We validated the effect on cell growth for two out of three E-boxes. For chr10_BS212 we did not observe decreased proliferation with dCas9, despite a similar effect on expression of two adjacent genes *PRKCQ* and *PRKCQ-AS1* with both WT Cas9 and dCas9. Since we showed that knockdown of *PRKCQ-AS1* decreased K562 cell growth, it is difficult to reconcile those contradictory results. This E-box is not localized in an amplified region, therefore the discrepancy between WT Cas9 and dCas9 cannot be attributed to massive DNA damage caused by the former. Similar inconsistencies have also been reported for targeting p53 binding sites with WT Cas9 and dCas9 with an overall limited overlap[39]. Differences between WT Cas9 and dCas9 observed when studying TF binding sites require further investigation.

Interestingly, our study provided also some novel insights into the grammar of E-boxes and their surrounding sequences. We observed significantly different frequencies of specific nucleotides at certain positions in essential vs non-essential E-boxes. This observation requires further study in other cell types but already provides novel indications about E-box functionality and demonstrates the usefulness of high-throughput CRISPR/Cas9 mutagenesis for studying TF binding sites. Previous studies showed that DNA flexibility and structure determined by the flanking sequences impacts binding of TFs[48]. For example, it was demonstrated that nucleotides directly flanking binding sites for the glucocorticoid receptor (GR) differed between strongly and weakly activated genes. Of note, motifs

flanked by G/C nucleotides showed the weakest activation[49]. Similarly, we observed a significant underrepresentation of motifs flanked by G/C among the essential E-boxes. Moreover, bioinformatics analysis of TF binding sites demonstrated that the DNA shape determined by flanking sequences impacts binding specificity of closely related TFs from the bHLH family: MYC, MAX and MAD[50].

With our approach we were able to pinpoint MYC-dependent genes essential for cancer cell growth. This included among others multiple novel genes such as: *RPL21*, *NOP9*, and *MED16*, as well as validated MYC targets (e.g. *PFAS, PTPN23*, and *RPLP2*), which confirms the accuracy of our approach. Importantly, using our strategy it is possible to determine which E-boxes are essential for cell viability and identify relevant regulated target genes. This is important since we demonstrated that not all E-boxes in a binding site affected cell growth and target gene expression. Similarly, not all genes adjacent to an E-box responded to E-box disruption and were crucial for cancer cells. In this study, we checked the expression of selected putative target genes with RT-qPCR post-screen. Combination of the MYC-CRISPR screen with single cell RNA sequencing would add another layer of information to this experimental setup and allow direct identification of genes responding to E-box disruption.

Genes localized near essential E-boxes were involved in various biological processes associated with tumorigenesis and overlapped with MYC-dependent pathways. Those included ribosome biogenesis, translation, RNA/DNA biosynthesis, intracellular transport and other processes crucial for cell proliferation and survival. In our study, we had a closer look at the purine biosynthesis pathway. Although all six enzymes in this pathway are known MYC targets (but it is not clear whether direct targets) and knockout of five of them strongly reduced K562 cell growth, only E-box in *PFAS* was essential in our screen. Disruption of E-boxes near other genes in this pathway affected their expression to a lesser extent or had no effect at all. This indicates that *PFAS* might be the critical MYC-regulated player in the purine biosynthesis pathway, although we cannot exclude the possibility that despite efficient E-box disruption for other genes in the pathway it was insufficient to diminish MYC binding. Similar pattern emerged for two other processes: splicing and mitochondrial fatty acid synthesis. In all cases our screen pinpointed the crucial MYC-dependent hubs, which may have implications for potential therapeutic approaches.

A limitation of our study is that we were not able to target all E-boxes within the MYC-bound loci. This was due to either lack of PAM sequence nearby or strong off-target activity of designed sgRNAs. While the second reason was not a major issue (only 4% of E-boxes were not targeted due to potential off-targets), because of lack of PAM sequence close to the E-box, 35% of E-boxes were not included in our library. This could be overcome with a complementary approach using variant Cas9 nucleases with different PAM requirements. In addition, dCas9 could offer more flexibility. In our

approach we required presence of PAM up to 3 nt upstream or downstream of the E-box, so that the cut occurs within the motif. With dCas9, PAM could be located further away from the E-box, as long as the sgRNA binding site covers the motif.

A potential flaw in our approach could be the fact that other bHLH proteins can bind to E-boxes and affect transcription such as: hypoxia-induced factor 1 (HIF1)[51], Twist2[52], Hey proteins[53], Zinc finger E-box–binding homeobox 1 (ZEB1) and 2 (ZEB2)[54] and MAX-binding MYC antagonists: MXI1, MAD proteins, MNT[55]. It is thus possible that the effect of E-box disruption might be also related to other interactors. However, several findings strongly suggest that MYC is involved. Firstly, we focused only on validated MYC binding sites (based on available MYC ChIP-seq data). Secondly, luciferase reporter assay for selected essential E-boxes showed that transcription was decreased after MYC knockout. Furthermore, ChIP performed for selected E-boxes confirmed that E-box disruption reduced MYC binding. Altogether, this indicates that although we cannot exclude that other factors may also be involved, MYC binding is crucial for the activity of the studied E-boxes.

In summary, the combined high-throughput screening assay using the MYC-CRISPR library targeting E-boxes and the Brunello library for gene knockout is a first powerful tool for genome-wide identification of MYC-dependent vulnerabilities in cancer cells. This well-validated novel approach allows for the identification of essential MYC binding sites and the regulated target genes which can serve as a basis for novel therapeutic approaches in MYC-dependent cancers. The broad design enables studies in a variety of normal and cancer cell types and determination of common as well as cell-type specific targets.

MATERIALS AND METHODS

Cell lines

The chronic myeloid leukemia cell line K562 (ordered from DSMZ, Germany) was maintained in Roswell Park Memorial Institute 1640 medium (Lonza, Switzerland) supplemented with 10% fetal bovine serum (Sigma-Aldrich, Saint Louis, MO, US) 2mM L-glutamine (Biowest, Nuaille, France), and 1% penicillin/streptomycin (Biowest) in a 5% CO_2 incubator at 37°C. HEK293T cell line (DSMZ) used for lentiviral particles production was maintained in low glucose Dulbecco's Modified Eagle's Medium (Lonza) supplemented as described above. Mycoplasma tests were routinely performed and confirmed that the cells were not contaminated.

Plasmids

The lentiCRISPR v2 (#52961) and lentiCRISPR v2-dCas9 (#112233) vectors were purchased from Addgene (Watertown, MA, US). The plasmids contains the coding sequence of *S. pyogenes* (d)Cas9 and

a cloning site for the sgRNA, as well as a puromycin resistance gene allowing selection of transduced cells. pGreenpuro lentivectors (SBI, Mountain View, CA, USA) with MYC shRNAs and scrambled control, a kind gift of prof. Anke van den Berg[56], were used for experiments with MYC knockdown.

Design of the MYC-CRISPR library

To design a library of single guide RNA (sgRNA) for comprehensive genome-wide targeting of MYC binding sites in cancer cells, we utilized publicly available MYC-ChIP-seq data from MCF7, K562 and HepG2 cells[32] and Burkitt lymphoma cell lines[20]. Genomic coordinates (hg19) of MYC-ChIP peaks were obtained using UCSC Table Browser and from the published data[20]. Genomic intervals were concatenated, merged, and DNA sequences were retrieved using Galaxy[57]. To avoid targeting coding exons, which could have resulted in disruption of the protein apart from the E-box, sequences overlapping with coding exons (based on Ensembl hg19 annotation) were removed. In the resulting sequences, E-box motifs were identified (canonical CACGTG, non-canonical CACATG/CATGTG and CACGCG/CGCGTG) and all sgRNAs targeting these E-boxes were designed based on the presence of the PAM sequence (NGG or CCN) using an in-house Python script (https://github.com/tomaszwozniakihg/cas9_search_tool). The resulting sgRNAs were checked for off-target binding using the CAS-OT script[58]. Only sgRNAs with at least three mismatches to the potential off-targets or at least two mismatches including at least one in the seed region (nt 9-20) were retained. The library also included 1,000 non-targeting sgRNAs as a negative control, and four sgRNAs against MYC as a positive control, all from the Brunello library. A list of all sgRNA oligonucleotides is provided in Table S1. The efficiency of sgRNAs was calculated with the WU-CRISPR tool[34].

Cloning of the MYC-CRISPR library and amplification of the Brunello library

Oligonucleotides containing the 20 nt sgRNA sequences flanked by the sequence from the lentiCRISPR_v2 vector TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCG [20 nt sgRNA] GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGT were synthesized by Twist Bioscience (San Francisco, CA, US). 2 ng of oligonucleotide pool was amplified with the oligo-F and oligo-R primers (Supplementary Table 8) using the NEBNext High-Fidelity Master Mix (New England Biolabs, Ipswich, MA, US) in 20 x 25 µl PCR reactions. PCR program: 98°C 30 sec; (98°C 10 sec; 63°C 10 sec; 72°C 15 sec) x 7 cycles; 72°C 2 min. PCR product was purified using QIAquick PCR Purification Kit (Qiagen, Hilden, Germany) and then extracted from an agarose gel with QIAquick Gel Extraction Kit (Qiagen). LentiCRISPRv2_puro vector was digested with the BsmBI restriction enzyme (New England Biolabs) and purified from an agarose gel. sgRNA MYC-CRISPR library was cloned into lentiCRISPRv2_puro vector using the circular polymerase extension cloning (CPEC) method as described previously[59]. Briefly, 20 CPEC reactions were performed, each using 100 ng digested vector, 10.8 ng amplified oligos and

NEBNext High-Fidelity Master Mix. PCR program: 98°C 30 sec; (98°C 10 sec; 72°C 7 min) x 5 cycles; 72°C 5 min. All PCR reactions were pooled and purified by isopropanol precipitation. 300 ng of the CPEC product was used for transformation of electrocompetent Endura cells (Lucigen, Middleton, WI, USA) according to the manufacturer's protocol. Fourteen electroporations were performed giving in total ~7.9 million colonies and resulting in ~170x coverage of the library. Bacteria were spread on 245x245 agar plates and grown for 14 h at 37°C. Colonies were scraped off the plates and plasmid DNA was isolated using Plasmid Plus Maxi Kit (Qiagen). The MYC-CRISPR library was deposited in Addgene (#173195). Brunello library targeting all human protein-coding genes[33] was purchased from Addgene (#73179). 50 ng of the library was electroporated into Endura cells. Four electroporations were performed resulting in >2,500x coverage. Quality of the MYC-CRISPR and Brunello libraries was verified by next-generation sequencing on Illumina platform (BGI, Hong-Kong).

Cloning of individual sgRNA constructs

For cloning of individual sgRNAs (Supplementary Table 9) into the lentiCRISPRv2_puro and lentiCRISPRv2-dCas9 vectors, sense and antisense oligos containing overhangs compatible with BsmBI sticky ends were synthesized by Genomed (Warsaw, Poland). Oligonucleotides were diluted in annealing buffer (10mM Tris-HCl pH8, 1mM EDTA pH8, 50mM NaCl). Annealing was performed in a thermocycler under conditions: 95°C 5min, 95°C (-1°C/cycle) x 70 cycles. Annealed sgRNAs were ligated into the lentiCRISPRv2 and lentiCRISPRv2-dCas9 vectors digested with BsmBI restriction enzyme at 1:5 vector:insert molar ratio with T4 DNA ligase (Invitrogen, Carlsbad, CA, US). 1 µl of ligation reaction was transformed into JM109 competent cells (Promega, Madison, WI, USA). Plasmid DNA from single colonies was isolated using Plasmid Plus Maxi Kit (Qiagen). Sequences of individual sgRNA constructs were confirmed by Sanger sequencing (Genomed).

Generation of lentiviral particles

For large scale production of lentiviral particles ~7.5 million HEK293T cells were plated in a T75 flask one day prior to transfection. Next day, ~80% confluent cells were transfected with packaging plasmids psPAX (11.2 µg) and pMD2.G (7.5 µg), and MYC-CRISPR or Brunello library plasmid (15 µg) using lipofectamine 2000 reagent (Invitrogen). One day after transfection medium was replaced with 7.5 ml DMEM supplemented with 10% FBS. Two and three days post-transfection Brunello and MYC-CRISPR lentiviral supernatants were filtered through 0.45 µm filter and stored at -80°C. For small scale production of lentiviral particles 1 million HEK293T cells were plated per well on a 6-well plate and transfected the next day with packaging plasmids psPAX (1.5 µg) and pMD2.G (1 µg), and lentiCRISPRv2 plasmid (2 µg) using calcium phosphate transfection method (Invitrogen). One day after transfection

medium was replaced with 1.1 ml DMEM supplemented with 10% FBS and lentiviral supernatant was collected as described above.

Determination of virus titer and transduction of K562 cells for screening

2.7 million K562 cells were plated per well in a 12-well plate and transduced with different amounts of virus. 4 mg/ml polybrene was added and cells were spun down in plates (33°C, 1000x g, 2h). After spinfection, additional 1 ml of medium was added. 24 h after transduction cells were washed and plated out in four wells for each condition, two wells with 3 µg/ml puromycin and two wells without puromycin. Medium with puromycin was changed after three days. After four days of selection, cells were counted and the percentage of surviving cells relative to cells not treated with puromycin was calculated. From this we determined the amount of virus resulting in ~30% surviving cells, which implies that approximately 85% of transduced cells were infected by a single virus.

78 million cells for the MYC-CRISPR library and 130 million cells for the Brunello library were transduced in duplicate with the amount of virus that results in ~30% transduced cells, in the same conditions as described above. After four days of selection with puromycin (T0) part of the cells was collected for DNA isolation. Remaining cells were further cultured for 20 population doublings. At each passage, the amount of cells corresponding to a 500x coverage (24 million for MYC-CRISPR, 38 million for Brunello library) were cultured in RPMI medium with 1 µg/ml puromycin and collected at the final timepoint (T1).

Preparation of libraries for next-generation sequencing of the plasmid pool or genomic DNA

sgRNA sequences from the MYC-CRISPR and Brunello library plasmids were amplified in PCR reaction using High Fidelity MasterMix 2x (New England Biolabs) and primers containing Illumina adaptors, an 8 nt barcode specific for each library (reverse primer) and a variable length (9-18 nt) stagger sequence to increase library complexity (forward primer)[60] (Supplementary Table 8). PCR products were pooled and purified using QIAquick PCR Purification Kit (Qiagen). Amplicons were analyzed on agarose gel and then extracted using QIAquick Gel Extraction Kit (Qiagen). Quality check and quantification of libraries were performed by qPCR using KAPA Library Quantification Kit (Roche, Basel, Switzerland).

Genomic DNA collected from K562 cells infected with MYC-CRISPR and Brunello libraries was isolated using GENTRA Puregene Kit (Qiagen). PCR was performed to amplify the sgRNA sequences integrated in genomic DNA. DNA from 23 million cells (MYC-CRISPR) and 38 million cells (Brunello) was amplified in 120 (MYC-CRISPR) and 190 (Brunello) individual 50 µl PCR reactions per sample with 3 µg DNA input as described above. PCR products were purified, analyzed on agarose gel, pooled based on band

intensities, and extracted from gel as described above. Determination of quality and quantification of libraries was performed by qPCR (Kapa Library Quantification Kit).

NGS and data analysis

NGS was performed on Illumina X-Ten platform at BGI (Hong-Kong). Reads were trimmed to remove adaptor sequences and split based on barcodes for individual samples. Number of reads obtained for each sample is given in Supplementary Table 10. For sgRNA enumeration, raw reads were processed with a Python script[60]. Only sgRNAs with no mismatches were counted. sgRNA counts were then used for DeSeq2 analysis with the CRISPRAnalyzeR tool (http://www.crispr-analyzer.org)[61]. Adjusted p-value 0.001 was used as a cut-off for identification of significantly depleted or enriched E-boxes (MYC-CRISPR library) and genes (Brunello library).

Determination of CRISPR/Cas9 editing efficiency by TIDE

To confirm CRISPR/Cas9 disruption of selected E-boxes, DNA was isolated from K562 cells at day 7 after transduction with individual sgRNAs targeting the selected E-boxes. Genomic regions of ~500-800 bp flanking E-box sequences were amplified by PCR (primer sequences in Supplementary Table 8). Amplicons were sequenced using the Sanger method (Genomed) and analyzed with TIDE calculator (https://tide-calculator.nki.nl)[62], using indel size range of 50.

RNA isolation and qRT-PCR

Total RNA was isolated from K562 cells using Quick-RNA™ Miniprep Kit (Zymo Research, Irvine, CA, US). cDNA was synthesized from 500 ng RNA using QuantiTect® Reverse Transcription Kit (Qiagen). qPCR on 5 ng cDNA was performed on CFX96 Touch qPCR System (BioRad, Hercules, CA, US) using PowerUp SYBR Green Master Mix (Applied Biosystems, Waltham, MA US). Expression was normalized to TBP. All experiments were conducted in two independent biological replicates, each with three technical replicates.

Growth assay

Growth assay was performed using CellTiter-Glo® Luminescent Cell Viability Assay (Promega). $1*10^3$ K562 cells infected with sgRNA lentiviral constructs were plated in triplicate on 96-well plates. 100 ul Cell-Titer Glo reagent diluted 1:2 in PBS was added per well after 1h (baseline level), 48h and 96h. The luminescent signal was measured using a GloMax microplate reader (Promega). Experiments were performed in three independent biological replicates. Growth rate was calculated at 48h and 96h relative to the 1h measurement.

Chromatin immunoprecipitation assay

10 M of K562 cells infected with sgRNAs targeting selected E-box sequences were fixed to crosslink DNA with chromatin-associated proteins according to the Active Motif protocol. Briefly, cells were fixed for 20 min by adding 1/10 volume of 37% formaldehyde solution and then neutralized by adding 1/20 volume of 2.5 M glycine. Next, K562 cells were washed using PBS-Igepal, snap-frozen on dry ice and stored at -80°C. Chromatin immunoprecipitation with anti-MYC antibody (sc-764, Santa Cruz, Dallas, TX, US) followed by qPCR was performed by Active Motif (La Hulpe, Belgium).

Luciferase reporter assay

To confirm that sequences containing selected E-boxes drive transcription in a MYC-dependent way, we conducted luciferase reporter assay. MYC binding sites as defined by peaks from MYC-ChIP-seq data were amplified by PCR from DNA of K562 cells. For chr11_BS79 the amplified sequence was shorter to omit two other non-essential E-boxes present in the peak. Primers contained sequences to create overhangs for XhoI and SacI restriction sites (Supplementary Table 8) for cloning into the pGL4.23 vector, upstream of the firefly luciferase gene under minimal promoter (#E8411, Promega). K562 cells were transduced with MYC shRNAs or scrambled control. After selection with puromycin for 4 days, $1x10^5$ cells were co-transfected using Lipofectamine LTX with 500 ng pGL4.23 vector with E-box constructs and 5 ng pRL-SV40 vector containing Renilla luciferase (#E2231, Promega). 24h after transfection cells were lysed and luminescence was measured using Dual Luciferase Reporter Assay on GloMax (Promega). Firefly luminescence was normalized to Renilla. The experiments were performed in three independent biological replicates, each with triplicate transfection.

Gene Ontology and Gene Set Enrichment Analysis

For each E-box, the nearest genes with transcription start site (TSS) within 50kb both upstream and downstream were retrieved using Galaxy. Gene ontology analysis was conducted using DAVID Functional Annotation Tool v6.8[63,64] on essential genes from Brunello screen and essential genes within 50kb of at least two-fold depleted E-boxes from MYC-CRISPR screen. Pre-ranked gene set enrichment analysis (GSEA)[65,66] was performed on log2 fold change values of all genes in the Brunello library and of genes near at least two-fold depleted E-boxes. Hallmark (H) and curated (C2) gene sets v7.2 were used for analysis.

Analysis of features associated with essential E-boxes

MYC-ChIP, H3K4me1, H3K4me3, H3K27ac and DNase HS data for K562 cells were obtained from Encode via UCSC Table Browser. Data were overlapped with E-box positions using Galaxy. Gene expression data for K562 were obtained from The Cancer Cell Line Encyclopedia[67].

Establishment of monoclonal cell lines

K562 cells were infected with sgRNA targeting the E-box on chromosome 10 (chr10_BS212) or a non-targeting sgRNA. After puromycin selection, cells were diluted in MethoCult™ H4434 Classic medium (Stemcell Technologies, Vancouver, Canada) to obtain on average 2 cells per well in a 96-well plate. After expansion, cells were harvested for DNA and RNA. Mutations were confirmed by Sanger sequencing (Genomed).

Statistical analysis

To determine the statistical significance of results from CellTiter-Glo, RT-qPCR and luciferase assay Student's t-test was applied. Enrichment of essential genes near essential E-boxes was assessed by the chi-square goodness of fit test. Cumulative distribution of predicted sgRNA efficiency scores for essential and non-essential E-boxes and genes was compared using the Kolmogorov-Smirnoff test. Correlations between histone marks were determined with the Pearson r^2 coefficient. All statistical analyses were performed in GraphPad Prism, p value <0.05 was considered significant.

ACKNOWLEDGEMENTS

This research was funded by the National Science Centre, Poland (grant no. 2016/23/D/NZ1/01611). AD-K, NR and AvdB were supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 952304.

We would like to thank Dr. Jeroen Guikema from University Medical Center Amsterdam, the Netherlands, for inspiring discussions.

Figures: 2A, 2C, 4C, 4E and 4F were created using BioRender.

AUTHORS CONTRIBUTIONS

AD-K planned and supervised the project and acquired funding; MK, MŻ, MP, MEK, WS, WŁ, IZ-S and AD-K performed the experiments; TW designed the MYC-CRISPR library sgRNAs; data were analyzed by MK, MP, WŁ, NR and AD-K; NR, JK and AvdB contributed to project conceptualization; MK and AD-K prepared the figures and wrote the manuscript. All authors read and approved the manuscript.

REFERENCES

1. Dang, C. V. MYC on the path to cancer. *Cell* **149**, 22–35 (2012).
2. Chen, H., Liu, H. & Qing, G. Targeting oncogenic Myc as a strategy for cancer treatment. *Signal Transduct. Target. Ther.* **3**, 1–7 (2018).
3. Lancho, O. & Herranz, D. The MYC Enhancer-ome: Long-Range Trancriptional Regulation of MYC in Cancer. *Trends Cancer* **4**, 810–822 (2018).

4. Bachmann, A. S. & Geerts, D. Polyamine synthesis as a target of MYC oncogenes. *J. Biol. Chem.* **293**, 18757–18769 (2018).
5. De Klein, A. *et al.* bcr Rearrangements and Translocation of the c-abl Oncogene in Philadelphia Positive Acute Lymphoblastic Leukemia. *Blood* **68**, 1369–1375 (2015).
6. Kalkat, M. *et al.* MYC deregulation in primary human cancers. *Genes (Basel).* **8**, 2–30 (2017).
7. Littler, S. *et al.* Oncogenic MYC amplifies mitotic perturbations. *Open Biol.* **9**, (2019).
8. García-Gutiérrez, L., Delgado, M. D. & León, J. Myc oncogene contributions to release of cell cycle brakes. *Genes (Basel).* **10**, (2019).
9. Stine, Z. E., Walton, Z. E., Altman, B. J., Hsieh, A. L. & Dang, C. V. MYC, metabolism, and cancer. *Cancer Discovery* vol. 5 1024–1039 (2015).
10. Arvanitis, C. & Felsher, D. W. Conditional transgenic models define how MYC initiates and maintains tumorigenesis. *Semin. Cancer Biol.* **16**, 313–317 (2006).
11. Wolf, E. & Eilers, M. Targeting MYC Proteins for Tumor Therapy. *Annu. Rev. Cancer Biol.* **4**, 61–75 (2020).
12. Herkert, B. & Eilers, M. Transcriptional repression: The dark side of Myc. *Genes and Cancer* **1**, 580–586 (2010).
13. Fan, J. S. *et al.* Time-dependent c-Myc transactomes mapped by array-based nuclear run-on reveal transcriptional modules in human B cells. *PLoS One* **5**, (2010).
14. Menssen, A. & Hermeking, H. Characterization of the c-MYC-regulated transcriptome by SAGE: Identification and analysis of c-MYC target genes. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 6274–6279 (2002).
15. Schuhmacher, M. *et al.* The transcriptional program of a human B cell line in response to Myc. *Nucleic Acids Res.* **29**, 397–406 (2001).
16. Yap, C. S., Peterson, A. L., Castellani, G., Sedivy, J. M. & Neretti, N. Kinetic profiling of the c-Myc transcriptome and bioinformatic analysis of repressed gene promoters. *Cell Cycle* **10**, 2184–2196 (2011).
17. Sabò, A. *et al.* Selective transcriptional regulation by Myc in cellular growth control and lymphomagenesis. *Nature* **511**, 488–492 (2014).
18. Kim, J., Lee, J. H. & Iyer, V. R. Global identification of Myc target genes reveals its direct role in mitochondrial biogenesis and its E-box usage in vivo. *PLoS One* **3**, (2008).
19. Krepelova, A., Neri, F., Maldotti, M., Rapelli, S. & Oliviero, S. Myc and Max genome-wide binding sites analysis links the Myc regulatory network with the polycomb and the core pluripotency networks in mouse embryonic stem cells. *PLoS One* **9**, 1–12 (2014).
20. Seitz, V. *et al.* Deep sequencing of MYC DNA-Binding sites in Burkitt lymphoma. *PLoS One* **6**, (2011).

21. Zeller, K. I. *et al.* Global mapping of c-Myc binding sites and target gene networks in human B cells. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 17834–17839 (2006).
22. Lin, C. Y. *et al.* Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**, 56–67 (2012).
23. Fernandez, P. C. *et al.* Genomic targets of the human c-Myc protein. *Genes Dev.* **17**, 1115–1129 (2003).
24. Marinkovic, D. *et al.* Identification of novel Myc target genes with a potential role in lymphomagenesis. *Nucleic Acids Res.* **32**, 5368–5378 (2004).
25. Cermelli, S., Jang, I. S., Bernard, B. & Grandori, C. Synthetic lethal screens as a means to understand and treat MYC-driven cancers. *Cold Spring Harb. Perspect. Med.* **4**, (2014).
26. Kessler, J. D. *et al.* A sumoylation-dependent transcriptional subprogram is required for Myc-driven tumorigenesis. *Science (80-.).* **335**, 348–353 (2012).
27. Toyoshima, M. *et al.* Functional genomics identifies therapeutic targets for MYC-driven cancer. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 9545–9550 (2012).
28. Zhou, B. *et al.* Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res.* **29**, 472–484 (2019).
29. Zhou, Z. *et al.* Identification of synthetic lethality of PRKDC in MYC-dependent human cancers by pooled shRNA screening. *BMC Cancer* **14**, 1–13 (2014).
30. Malempati, S. *et al.* Aberrant stabilization of c-Myc protein in some lymphoblastic leukemias. *Leukemia* **20**, 1572–1581 (2006).
31. Huang, H. L. *et al.* Triggering Fbw7-mediated proteasomal degradation of c-Myc by oridonin induces cell growth inhibition and apoptosis. *Mol. Cancer Ther.* **11**, 1155–1165 (2012).
32. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
33. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
34. Wong, N., Liu, W. & Wang, X. WU-CRISPR: Characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biol.* **16**, 1–8 (2015).
35. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science (80-.).* **350**, 1096–1101 (2015).
36. Pedley, A. M. & Benkovic, S. J. A New View into the Regulation of Purine Metabolism: The Purinosome. *Trends Biochem. Sci.* **42**, 141–154 (2017).
37. Viola, I. L. & Gonzalez, D. H. *Methods to Study Transcription Factor Structure and Function*. *Methods to Study Transcription Factor Structure and Function. Plant Transcription Factors: Evolutionary, Structural and Functional Aspects* (Elsevier Inc., 2016). doi:10.1016/B978-0-12-

800854-6.00002-6.

38. Hartl, M. The quest for targets executing MYC-dependent cell transformation. *Front. Oncol.* **6**, 1–9 (2016).

39. Borys, S. M. & Younger, S. T. Identification of functional regulatory elements in the human genome using pooled CRISPR screens. *BMC Genomics* **21**, 1–15 (2020).

40. Han, R. *et al.* Functional CRISPR screen identifies AP1-associated enhancer regulating FOXF1 to modulate oncogene-induced senescence. *Genome Biol.* **19**, 1–13 (2018).

41. Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192–197 (2015).

42. Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat. Biotechnol.* **34**, 192–198 (2016).

43. Kim, Y. W. & Kim, A. R. *Deletion of transcription factor binding motifs using the CRISPR/spCas9 system in the β-globin LCR*. *Bioscience Reports* vol. 37 (2017).

44. Durbin, A. D. *et al.* Selective gene dependencies in MYCN-amplified neuroblastoma include the core transcriptional regulatory circuitry. *Nat. Genet.* **50**, 1240–1246 (2018).

45. Dominguez, A. A., Lim, W. A. & Qi, L. S. Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation. *Nat Rev Mol Cell Biol.* **17**, 5–15 (2016).

46. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens Molly. *Cell* **176**, 377–390 (2019).

47. Klann, T. S. *et al.* CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat. Biotechnol.* **35**, 561–568 (2017).

48. Yella, V. R. *et al.* Flexibility and structure of flanking DNA impact transcription factor affinity for its core motif. *Nucleic Acids Res.* **46**, 11883–11897 (2018).

49. Schöne, S. *et al.* Sequences flanking the core-binding site modulate glucocorticoid receptor structure and activity. *Nat. Commun.* **7**, (2016).

50. Yang, L. *et al.* TFBSshape: A motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res.* **42**, D148-155 (2014).

51. Schmitz, G. & Langmann, T. Transcriptional regulatory networks in lipid metabolism control ABCA1 expression. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids* **1735**, 1–19 (2005).

52. Chengxiao, Z. & Ze, Y. Biological function and molecular mechanism of Twist2. *Yi Chuan* **37**, 17–24 (2015).

53. Weber, D., Wiese, C. & Gessler, M. *Hey bHLH transcription factors*. *Current Topics in Developmental Biology* vol. 110 (Elsevier Inc., 2014).

54. Jiang, Y. *et al.* Zinc finger E-box– binding homeobox 1 (ZEB1) is required for neural differentiation of human embryonic stem cells. *J. Biol. Chem.* **293**, 19317–19329 (2018).

55. Hurlin, P. J. & Huang, J. The MAX-interacting transcription factor network. *Semin. Cancer Biol.* **16**, 265–274 (2006).

56. Winkle, M. *et al.* Long noncoding RNAs as a novel component of the Myc transcriptional network. *FASEB J.* **29**, 2338–2346 (2015).

57. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44**, W3–W10 (2016).

58. Xiao, A. *et al.* CasOT : a genome-wide Cas9 / gRNA off-target searching tool. *Bioinformatics* **30**, 1180–1182 (2014).

59. Quan, J. & Tian, J. Circular polymerase extension cloning of complex gene libraries and pathways. *PLoS One* **4**, (2009).

60. Joung, J. *et al.* Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. *Nat. Protoc.* **12**, 828–863 (2017).

61. Winter, J. *et al.* CRISPRAnalyzeR: Interactive analysis, annotation and documentation of pooled CRISPR screens. *bioRxiv* (2017) doi:10.1101/109967.

62. Brinkman, E. K., Chen, T., Amendola, M. & Van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42**, (2014).

63. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).

64. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).

65. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

66. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).

67. Barretina, J. *et al.* Erratum: Addendum: The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity (Nature (2012) 483 7391 (603-607)). *Nature* **565**, 603–607 (2019).

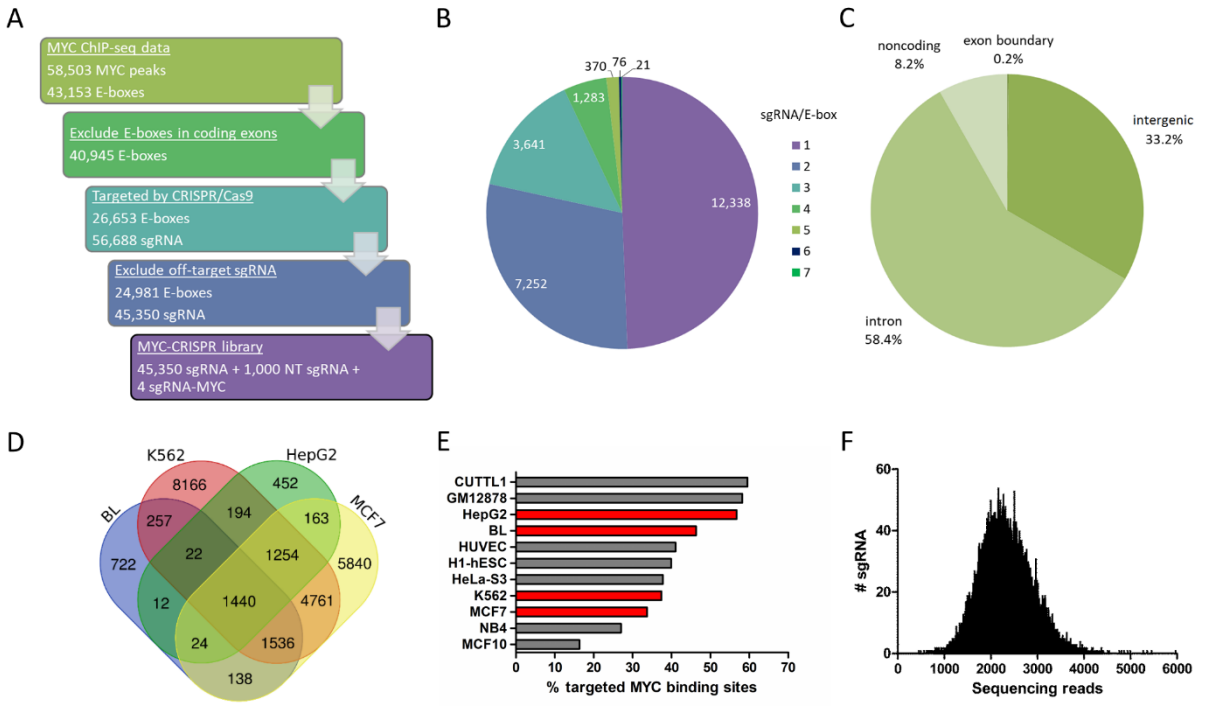
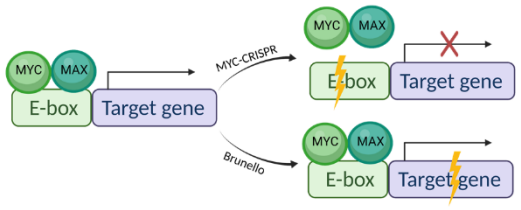


Figure 1. Design and generation of the MYC-CRISPR library for genome-wide disruption of MYC binding sites. A) Library was designed based on publicly available MYC-ChIP-seq data in MYC-dependent K562, MCF7, HepG2 and Burkitt lymphoma (BL) cell lines. After excluding E-boxes in coding exons, all possible sgRNAs targeting remaining E-boxes were designed, based on the presence of PAM sequence. sgRNAs with predicted off-target binding were filtered out. Final library contains 43,350 sgRNAs targeting E-boxes, 1,000 non-targeting (NT) sgRNAs as a negative control, and four sgRNAs targeting MYC as a positive control. **B)** Number of sgRNA constructs per E-box. **C)** Genomic location of E-boxes targeted by the library. **D)** Overlap of targeted E-boxes for selected cancer cell lines. **E)** Percentage of MYC binding sites targeted by the MYC-CRISPR library in various cel lines, based on available MYC-ChIP-Seq data. In red are cell lines for which the library was designed. **F)** Distribution of sgRNA constructs in the MYC-CRISPR plasmid library determined by NGS. All sgRNAs were present in the library.

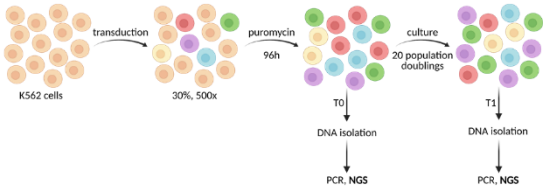
A



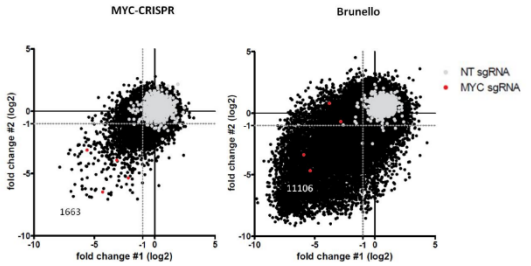
B

Library	% of transduction	Coverage
MYC-CRISPR #1	32	540x
Brunello #1	23	390x
MYC-CRISPR #2	26.9	455x
Brunello #2	24.5	415x

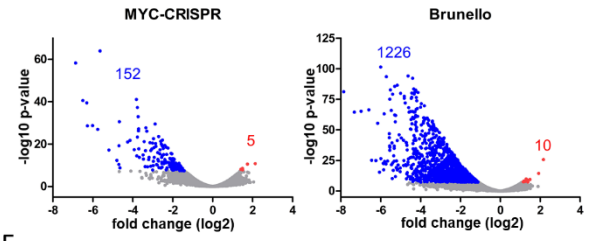
C



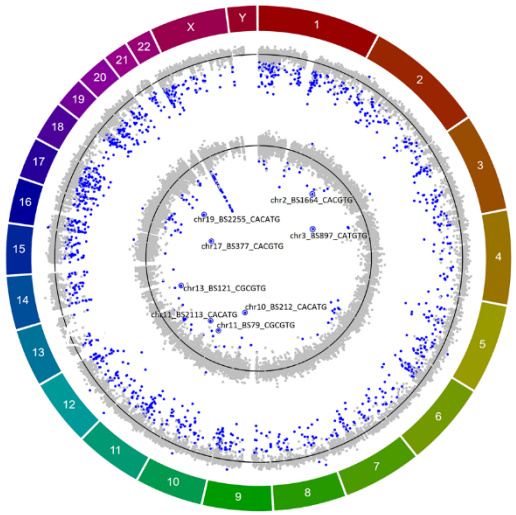
D



E



F



G

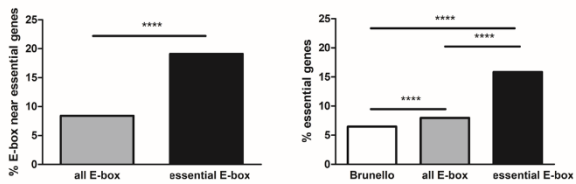


Figure 2. High-throughput screen with MYC-CRISPR and Brunello libraries in K562 cells. A) Experimental approach: 1) MYC-CRISPR library to destroy E-box sequences, disrupt MYC binding and its effect on target gene expression 2) Brunello library for genome-wide gene knock-out. B) Performance of the screen in K562 cells conducted in duplicate. C) Scheme of the high-throughput screen in K562 cells with MYC-CRISPR and Brunello libraries. D) Scatterplots showing changes in sgRNA abundance in two screen replicates. Dotted line indicates FC=2. None of the non-targeting sgRNAs was depleted ≥2-fold; 2-4 sgRNA targeting MYC were consistently depleted. 1,663 sgRNA from the MYC-CRISPR library and 11,106 sgRNA from the Brunello library were consistently ≥2-fold depleted in both replicates. E) DESeq2 analysis revealed 152 essential E-boxes and 1226 essential genes in K562, while 5 E-boxes and 10 genes were significantly enriched. F) Circos plot showing log2FC values for genes in Brunello screen (outer circle) and E-boxes in MYC-CRISPR screen (inner circle) across the chromosomes. Blue dots indicate genes and E-boxes significantly (padj<0.001) depleted or enriched, black lines denote log2FC=0. The most significant E-boxes are indicated. G) The percentage of E-boxes localized near essential genes among all and essential E-boxes (left). The percentage of essential genes identified in the Brunello library and localized near all and essential E-boxes (right).

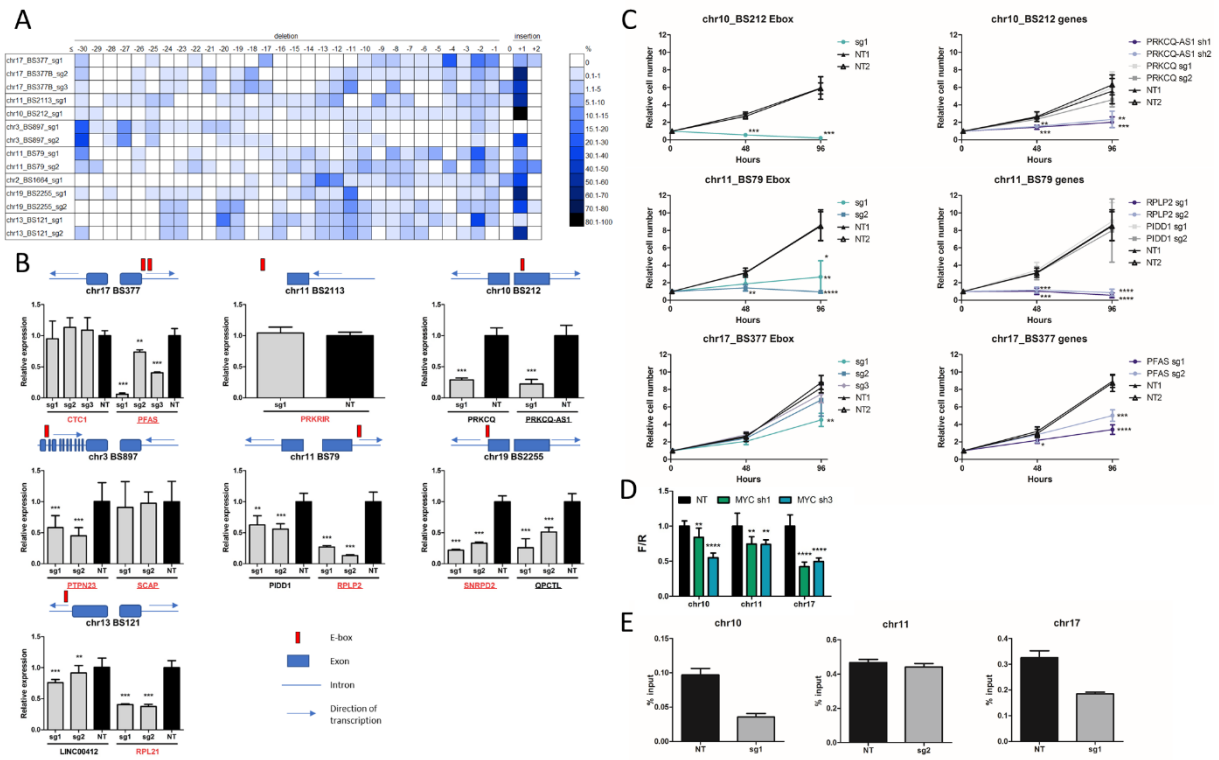


Figure 3. Validation of selected E-boxes and target genes. A) Efficiency of disruption of selected E-boxes and the spectrum of mutations introduced by individual sgRNAs, demonstrated by TIDE analysis. Size distribution of introduced indels ranged from ≤-30 bp to +2 bp. Colors indicate percentage of sequences with a given indel size. **B)** qRT-PCR analysis of genes adjacent to selected E-boxes upon CRISPR/Cas9 disruption of E-box sequences. For six out of seven E-boxes at least one nearby gene showed significantly decreased expression. Known MYC-regulated genes are underlined; genes essential or at least 4-fold depleted in Brunello screen are in red. Mean and SD of two independent experiments, each performed in triplicate, are shown. **, p<0.01; ***, p<0.001, Student's t-test. **C)** Cell viability upon disruption of selected E-boxes and knockout of adjacent genes was measured using CellTiter-Glo assay at three timepoints: 0, 48 and 96 hours. Shown are mean values and SD from three independent experiments, each performed in triplicate. *, p<0.05; **, p<0.01; ***, p<0.001; ****, p<0.0001, Student's t-test. **D)** Luciferase reporter assay for selected E-boxes upon MYC knockdown with shRNA. Decreased luminescence signal was observed for all E-boxes in MYC-shRNA samples vs. NT control. **, p<0.01; ****, p<0.0001, Student's t-test. Mean and SD of three independent experiments, each performed in triplicate, are shown. **E)** MYC ChIP-qPCR analysis of MYC binding upon E-box disruption. Cells were infected with sgRNAs targeting selected E-boxes. MYC binding was decreased for chr17_BS377 and for chr10_BS212 but not for chr11_BS79.

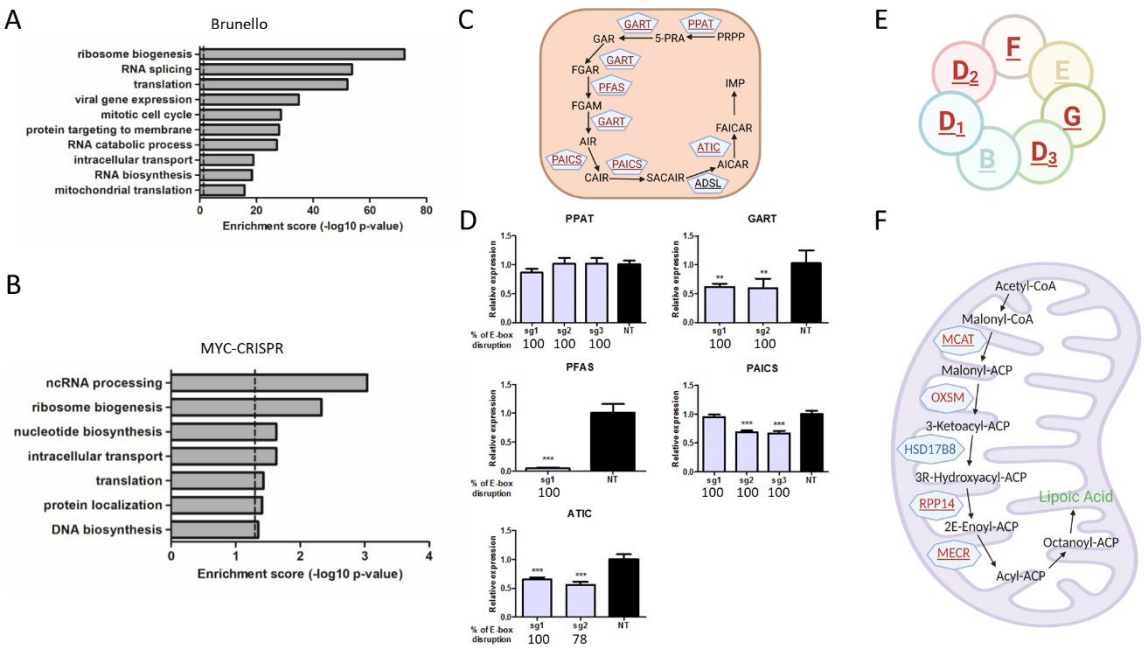


Figure 4. Identification of critical MYC-dependent hubs in pathways affected by MYC. Top 10 Gene Ontology clusters enriched among **A)** the essential genes (p_{adj}<0.001) identified in Brunello screen, and **B)** genes localized near at least 2-fold depleted E-boxes identified in MYC-CRISPR screen. Dotted line - p<0.05 **C)** Purine biosynthesis pathway scheme. Enzymes are shown in boxes, substrates/products in black font. **D)** Efficiency of E-box disruption and qRT-PCR analysis of genes involved in the purine biosynthesis pathway upon CRISPR/Cas9 disruption of E-box sequences. Mean and SD of two independent experiments, each performed in triplicate, are shown. **, p<0.01; ***, p<0.001, Student's t-test. **E)** Scheme of the spliceosomal U1 small nuclear ribonucleoprotein (snRNP) core, formed by 7 Sm proteins: B, D1, D2, D3, E, F and G. **F)** Scheme of the mitochondrial fatty acid synthesis pathway. Enzymes are shown in boxes, substrates in black font. **C, E and F)** Known MYC-regulated genes are underlined, genes essential or at least 4-fold depleted in Brunello screen are in red.

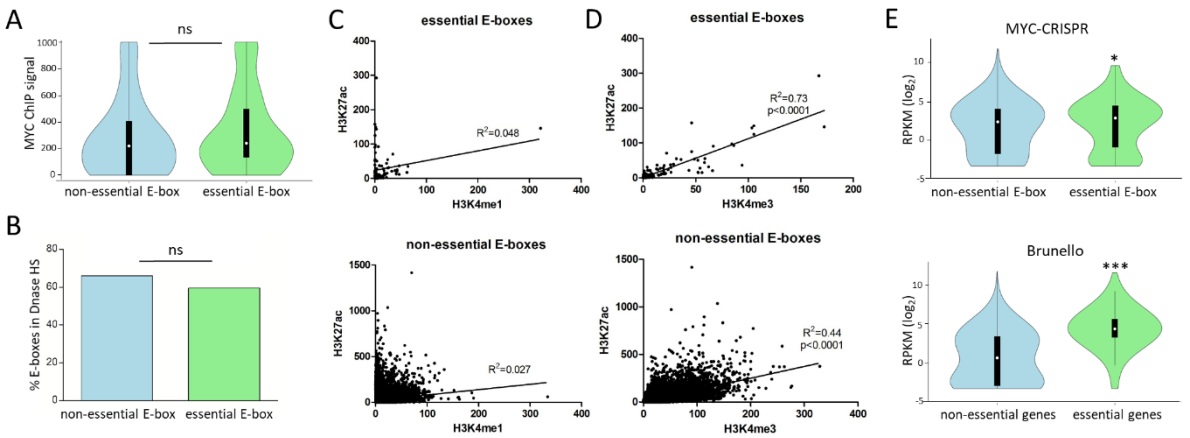


Figure 5. Features of essential E-boxes. A) MYC binding strength to essential and non-essential E-boxes based on MYC-ChIP signal. ns, Student's t-test. **B)** Localization of essential and non-essential E-boxes within DNase hypersensitivity regions. ns, chi-square test. **C)** Correlation between histone marks H3K27ac and H3K4me1 as a feature of active enhancers. **D)** Correlation between histone marks H3K27ac and H3K4me3 as a feature of active promoters. Pearson R^2 coefficient. **E)** Expression of adjacent genes localized near essential and non-essential E-boxes (upper graph) and expression of essential and non-essential genes from Brunello library (lower graph), *, $p<0.05$; ***, $p<0.001$ Student's t-test.

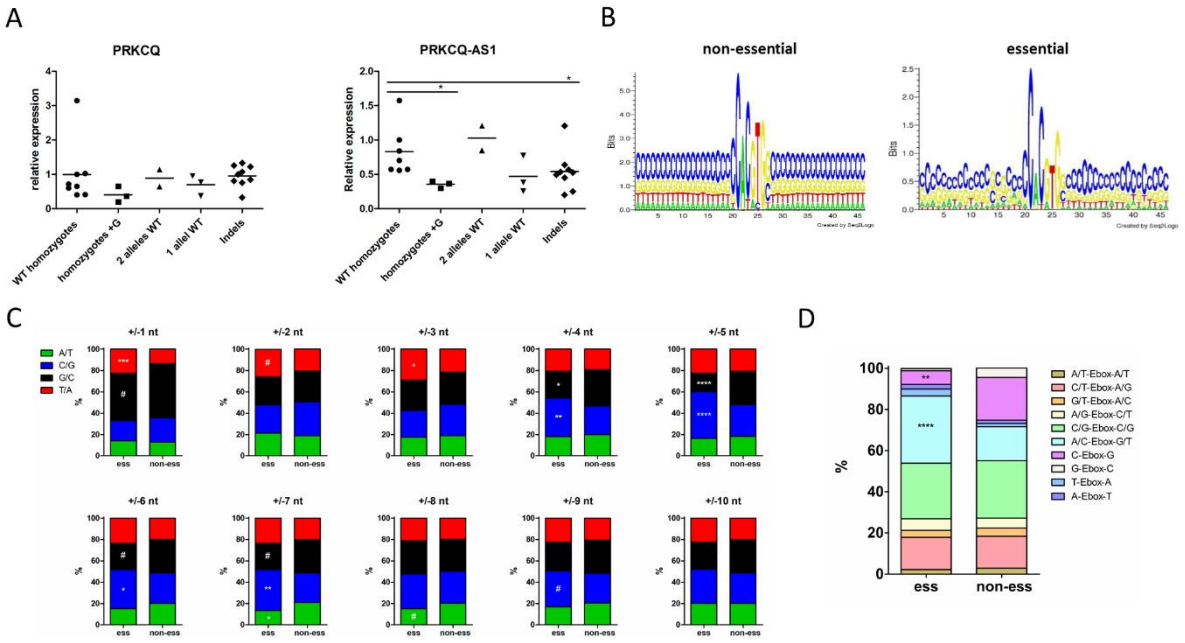


Figure 6. Grammar of the E-box sequence context. A) Expression of two genes adjacent to an E-box on chromosome 10 (chr10_BS212) was examined in monoclonal cell lines derived from K562 cells transduced with an sgRNA targeting this E-box. A spectrum of clones with various modifications of the E-box were obtained, including homozygotes with the +G insertion after E-box which seemingly did not change the E-box sequence. *, p<0.05, Kruskal-Wallis test with Dunn's post test. **B)** Sequence logo (created using Seq2Logo) of the E-boxes and 20 nt flanking sequences for non-essential and essential E-boxes. **C)** Frequency of up to 10 nt upstream/downstream flanking essential and non-essential E-boxes. Since E-boxes are (quasi)palindromic and can be read on either strand, G at +1 equals C at -1 etc. **D)** Frequency of pairs of two immediate adjacent nucleotides flanking essential and non-essential E-boxes. For C) and D) #, p<0.1; *, p<0.05; **, p<0.01; ***, p<0.001; ****, p<0.0001; chi-square goodness of fit test.