Computational prediction of CRISPR-impaired non-coding regulatory regions

Nina Baumgarten[1,2,3,*], Florian Schmidt[1,2,3,4,*], Martin Wegner[5,6], Marie Hebel[5,6], Manuel Kaulich[5,6], Marcel H. Schulz[1,2,3]

[1] Institute for Cardiovascular Regeneration, Goethe University, 60590 Frankfurt am Main, Germany,

[2] German Center for Cardiovascular Research, Partner site Rhein-Main, 60590 Frankfurt am Main, Germany,

[3] Cluster of Excellence MMCI, Saarland University, and Max Planck Institute for Informatics, Saarland Informatics Campus, 66123 Saarbrücken, Germany,

[4] Laboratory of Systems Biology and Data Analytics, Genome Institute of Singapore, 60 Biopolis Street, 138672 Singapore, Singapore,

[5] Institute of Biochemistry II, Goethe University - Medical Faculty, University Hospital, 60590 Frankfurt am Main, Germany,

[6] Frankfurt Cancer Institute, Goethe University, 60590 Frankfurt am Main, Germany,

∗ joint first authors,

Correspondence to: marcel.schulz@em.uni-frankfurt.de

Abstract. Genome-wide CRISPR screens are becoming more widespread and allow the simultaneous interrogation of thousands of genomic regions. Although recent progress has been made in the analysis of CRISPR screens, it is still an open problem how to interpret CRISPR mutations in non-coding regions of the genome. Most of the tools concentrate on the interpretation of mutations introduced in gene coding regions. We introduce a computational pipeline that uses epigenomic information about regulatory elements for the interpretation of CRISPR mutations in non-coding regions. We illustrate our approach on the analysis of a genome-wide CRISPR screen in hTERT-RPE-1 cells and reveal novel regulatory elements that mediate chemoresistance against doxorubicin in these cells. We infer links to established and to novel chemoresistance genes. Our approach is general and can be applied on any cell type and with different CRISPR enzymes.

Keywords: CRISPR-Cas9 screen, mutations, non-coding genome, gene regulation, chemoresistance

Introduction

One of the largest problems in current biology and molecular medicine is understanding how genotypes cause phenotypes. The recent development of the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) technology accelerated genotype to phenotype associations by introducing targeted

2

gene perturbations [1]. While arrayed CRISPR experiments enabled the association of phenotypes to the perturbation of individual genes, genome-wide CRISPR experiments are able to target entire genomes simultaneously in a single experiment [2,3]. Unbiased genome-wide CRISPR screens facilitated the linking of genotypes with phenotypes on a large scale and facilitated the identification of drug resistance and cancer vulnerabilities [4–7]. In addition to fishing for potential hit candidates on a genome-wide scale, the CRISPR technology has been used to understand gene architecture and identify functional protein domains through exhaustive gene tiling [8,9]. Here, all possible gRNAs for a locus of interest were tested simultaneously to identify gene/protein segments in which gRNAs were under/over-represented due to a phenotypic enrichment. Moreover, CRISPR screens have been applied to the non-coding genome to functionally characterize long non-coding RNAs (lncRNAs) and cis-regulatory elements linked to cell viability and chemotherapy resistance [3, 10, 11].

Current approaches to analyze genome-wide CRISPR screens are based on computing log fold changes of gRNA abundance between treated and untreated cell populations. Usually, the treated cell population is harvested at the end time point of the screen and compared to an untreated control population that has been harvested at an early time point, or even the plasmid library. Positive and negative hit gRNAs are then called based on the ranking of these relative changes [7, 12]. This principle can be applied to analyze CRISPR screens targeting protein-coding genes as well as non-coding regions. For CRISPR screens targeting known protein-coding genes, a variety of statistical analysis tools are available: One of the first algorithms that was specifically designed for the analysis of CRISPR knockout screens was MAGeCK, around which an entire suite of visualization and analysis tools has evolved [13–15]. The BAGEL algorithm aims to identify hitherto unknown essential genes by applying a statistical model that is based on prior knowledge about gene essentialities [16]. A user-friendly and easily accessible tool is PinAPL-Py which is implemented as a web service and offers multiple analysis options for a variety of CRISPR screening experiments [17]. However, there are no general approaches available to call hits of a genome-wide screen using a randomized gRNA library that targets mainly non-coding regions. The reason is that an additional annotation step is required to overlap hit gRNA target sites in non-coding regions with known or predicted gene regulatory elements. Also, it is currently unclear whether the same statistical assumptions hold true for calling hits in coding and non-coding regions of the genome.

Understanding the regulation through non-coding regions is a current topic in genetics. One particular role of the non-coding genome is to harbor Regulatory EleMents (REMs) such as enhancers, repressors, and promoters. REMs can be bound by transcription factors (TFs). TFs can recruit other proteins that can influence the 3D-structure of the chromatin and regulate gene expression in a cell-type specific manner [18]. REMs can be located far away from the genes they regulate and lead to up- or down-regulation of target gene expression [19–21]. Identifying REMs is challenging, and different genome-

wide protocols can be used. There is no one gold-standard method, but instead different (epi-)genomic indicators and different algorithmic approaches are used for the annotation of REMs [22–28].

There are a number of databases that describe human REMs, such as the *Vista Enhancer Browser* [29], with experimentally verified enhancers, the FANTOM5 enhancer resource [26], or the *HACER* database [30]. Furthermore, there are resources that hold information about REMs and their putative target genes, the genes they regulate. Examples are the GeneHancer database [31] or *RAEdb* [32]. Another large-scale source of human REMs is *EpiRegio* [33], a database that contains 2.4 million predicted REMs with associated target genes predicted with the STITCHIT algorithm [28]. Databases such as *HEDD* [34] or *DiseaseEnhancer* [35] list REMs that are associated to known disease genes, highlighting the role of non-coding variation in human disease.

In this study we develop and test a workflow for the prediction of non-coding CRISPR events that disrupt gene regulation based on the annotation of human REMs and epigenomic cell-specific information. We apply our approach to a genome-wide CRISPR screen that has been performed in hTERT RPE-1 cells to find coding and non-coding regions that mediate doxorubicin resistance [3].

Doxorubicin is a chemotherapy medication used to treat different forms of cancer. It induces double-strand DNA breaks and triggers DNA damage associated cell cycle arrest and apoptosis pathways, for example via MAPK/ERK pathway [36,37]. Our analysis reveals 35 genes that we could link to doxorubicin resistance through genomic pertubations in active REMs in RPE-1 cells.

Results

Overview of the suggested approach

Our approach identifies regulatory elements (REMs) linked to their target genes, which are (1) genomically modified by a gRNA and (2) are active in the cell type of interest. In order to identify the genomically modified REMs, we intersect them with the gRNAs binding sites. Additionally, epigenomic data measuring open chromatin, *e.g.* DNase1-seq, or histone modifications associated with active transcription, *e.g.* H3K27ac or H3K4me3, can be used to identify REMs active in the cell type of interest. Based on the active and genomically modified REMs, a protein-protein association network of the associated target genes is constructed and a motif enrichment analysis of the REMs is performed. A general overview of our approach is shown in Fig. 1.

Prediction of regulatory regions that mediate chemoresistance in RPE-1 cells

To illustrate our newly proposed approach on a realistic application, we applied it to a genome-wide doxorubicin CRISPR-Cas9 resistance screen in hTERT-RPE-1 cells by Wegner *et al.* [3]. As a result of the screen, 332 non-coding target sites of 226 validated gRNA sequences were identified. We extended
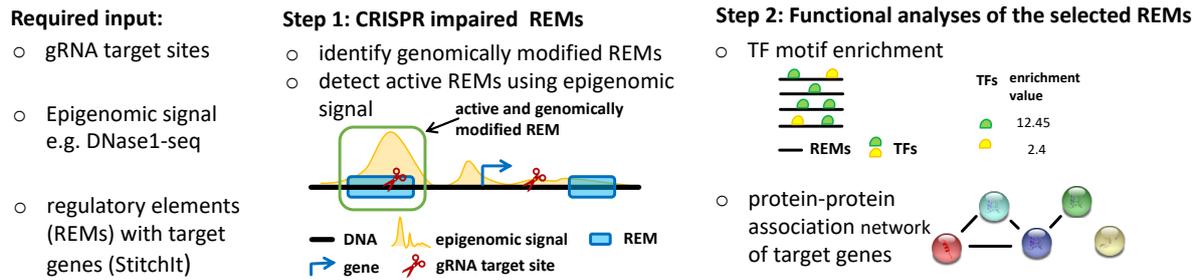
4

Fig. 1. Overview of our analysis approach. Step 1: gRNA target sites are overlapped with a catalogue of gene regulatory elements (REMs) and peak regions of epigenomic measurements for regulatory activity (*e.g.* DNase1-seq, H3K4me3). Step 2: functional analysis of the target genes reveals involved complexes. Transcription factor (TF) motif enrichment in REMs identifies TFs that are involved in the trait of interest.

the gRNA target sites to regions of length 200bp such that the target site is centered in the middle. We aim to identify REMs affected by gRNAs, and explore their possible involvement in chemoresistance.

The REMs and their target genes were inferred simultaneously by Schmidt *et al.* using STITCHIT [28]. This method interprets the variation in epigenetic data, like DNase1-seq in relation to gene expression, across various cell types. To take even distal REMs into account, STITCHIT 's predicted REMs are within a window covering 100kb upstream of the gene's transcription start site, the gene body and 100kb downstream of the gene's transcription termination site. In total, we downloaded 3,900,708 REMs associated to 36,817 genes based on data from the Blueprint [38], Roadmap [39] and the ENCODE [40] consortium.

We found 219 putative REMs overlapping with a non-coding gRNA, which could be linked to 190 different genes (see Sup. Table).

To ensure that we observe the overlap not only by chance, we randomly shuffled the positions of gRNA target sites within the genome and intersected the random regions with the REMs from our catalogue. We repeated this 100 times and obtained on average 72.07, 83.87 and 25.45 REMs overlapping with the sampled regions for Blueprint, Roadmap and ENCODE, STITCHIT REMs, respectively. We used a two-sided t-test to compare results from the shuffling analyses with the REMs overlapping true gRNAs and found a significant enrichment in each data set (Blueprint: 89/72.07, p-value $< 2.2e\text{-}16$; Roadmap: 90/83.87, p-value 1.638e-06; ENCODE: 40/25.45, p-value $< 2.2e\text{-}16$).

Analysis of epigenomic data reveals strong candidate doxorubicin chemoresistance genes

STITCHIT identifies REMs using epigenomics data from several cell types, therefore we can not directly conclude whether or not REMs affected by a gRNA are *active* in hTERT-RPE-1 cells. To identify active REMs in open chromatin regions, we have integrated DNase1-seq and H3K4me3 ChIP-seq data from human RPE-1 cells [40]. In total, we identified 13 gRNA target sites, which are overlapping active REMs linked to 35 different genes (see Table 1). Interestingly, it occurred several times that a gRNA target site

chr	gRNA coordinate	gene(s) associated
1	40,036,652	CAP1, PPT1, AL663070.1
10	103,282,011	**NT5C2** [41]
10	3,197,670	**PFKP** [42]
10	73,358,465	CFAP70, RPL26P6, **ANXA7** [43, 44], RNU6-833P
13	87,671,102	SLITRK5
16	30,021,011	**TMEM219** [45], **DOC2A** [46], TBX6, **GDPD3** [46, 47]
17	28,404,283	AC002094.1, KRT18P55, SEBOX, **TMEM199** [48]
17	42,423,866	MIR5010
17	81,635,498	MRPL12, **SLC25A10** [49, 50], AC137896.1, HGS, TSPAN10, AC139530.3, CCDC137
20	45,993,823	NCOA5, **MMP9** [51], PCIF1
22	24,063,890	AC253536.3
4	139,651,636	AC112236.2, H3P16, **MGST2** [52]
8	120,445,481	MTBP, COL14A1

Table 1. Candidate doxorubicin chemoresistance genes. All gRNAs that overlap regulatory regions with epigenetic evidence and their associated target genes are listed. Genes which have been reported in the literature to be linked to chemoresistance or cancer growth are in bold.

overlapped a region, which is linked to several target genes. These genes may be strong candidates to be involved in doxorubicin chemoresistance, as gRNA-mediated change of the DNA sequence may impair gene regulatory functions.

We have performed a detailed literature survey to find previous studies that have linked our candidate genes to chemoresistance against doxorubicin or other cancer drugs, supporting a putative role of REM target genes. We also considered studies that linked candidate genes to cancer growth, which may help to establish chemoresistance. In total, we have found that 8 out of 13 gRNAs link to at least one gene with literature evidence, as detailed below.

In myocytes, an increase of Myocardial Metallo Proteinases (MMPs) expression through increasing ROS formation induced by doxorubicin was observed [51]. At position chr20:45993823, overlap between a gRNA and REMs assume a regulatory site of MMP9, being in agreement with the finding of Spallarossa *et al.* [51]. Sun *et al.* showed that in colorectal cancer NCOA5 is upregulated, which leads to an higher expression of MMP9 and Cyclin D1 as well as to a lower expression of p27 through PI3K/AKT pathway. As a consequence, cell proliferation, migration and invasion are enhanced [53].

Another candidate is the NT5C2 gene, which is known to confer treatment resistance against thiorubin in acute lymphocytic leukemias [41]. For ANXA7 it was shown that resistance in different cancer types has been observed [43, 44].

One particularly interesting region was a gRNA binding to active REMs in a locus on chromosome 16. For three out of the four target genes we have found literature evidence. TMEM219 is a IGFBP-3

6

receptor that has been linked to promoting anti-tumoric effects of IGFBP-3 [45]. The gene expression and protein abundance of GDPD3 is associated with positive response to neoadjuvant chemotherapy in urothelial carcinomas [47]. A genome-wide screen for induced genes after doxorubicin in cancer cell lines [46] found that from the genes in Table 1 GDPD3 and DOC2A showed differential expression.

The expression of the TMEM199 gene was found to be upregulated in cisplatin resistent oesophageal adenocarcinoma cells pointing to a role in chemoresistance [48]. The gene product of SLC25A10 is a protein that belongs to the family of mitochondrial carriers, that are potential cancer therapy targets [49]. For example, high protein levels of SLC25A10 are associated with high metastasis potential and lower relapse-free survival in osteosarcomas [50]. Moon and others [54] found PFKP expression is upregulated in breast cancer cells. Further, high glycolytic metabolism, has been linked to breast cancer aggressiveness and growth [42].

Dvash *et al.* showed that endoplasmic reticulum stress and doxorubicin or 5-fluorouracil induce expression of MGTS2, and by that LTC4, which triggers DNA damage and cell death [52]. MGTS2-deficient mice seem to be resistance against the treatment with 5-fluorouracil. Further, they hypothesize that the MGST2-LTC4 pathway is not activated by commonly used cancer drugs, *e.g.* doxorubicin or 5-fluorouracil, in cells lines not able to express MGTS2, such as haematopoietic cell lines. They argue this could explain their chemoresistance.

As we found that many genes from Table 1 have been linked to chemoresistance, we wondered if they may form functional modules in the cell. We used the STRING database [55] to compute a protein-protein association network that contains all genes from Table 1 (see Fig. 2). The analysis revealed that DOC2A, TMEM219, GDPD3 and TBX6 form an association module. This is because all four genes lie within a region of 600kb, which is connected to different forms of human disease depending on whether the region is deleted or duplicated [56, 57].

Enrichment of transcription factors binding in chemoresistance regulatory elements

To identify key TFs involved in the regulatory process of doxorubicin resistance, we performed a TF motif enrichment analysis based on the active, genomically modified REMs. The analysis is done with *PASTAA* [58], where we have ranked sequences according to the open chromatin signal to detect enriched TFs compared to background REMs (more details in the methods section).

Table 2 lists the most enriched TFs and their corresponding FDR corrected p-value. The table contains a number of TFs that have been linked to chemoresistance. For example NR2C2 (or testicular nuclear receptor TR4) has been shown to suppress prostate cancer invasion by preventing infiltration of macrophages [59] and a recent report showed that the sensitivity to Docetaxel chemotherapy can be improved using antagonists against NR2C2 [60] in prostate cancer. Further, the TF NR2F2 (or Chicken ovalbumin upstream promoter transcription factor II, COUP-TF2) expressed in different types of col-
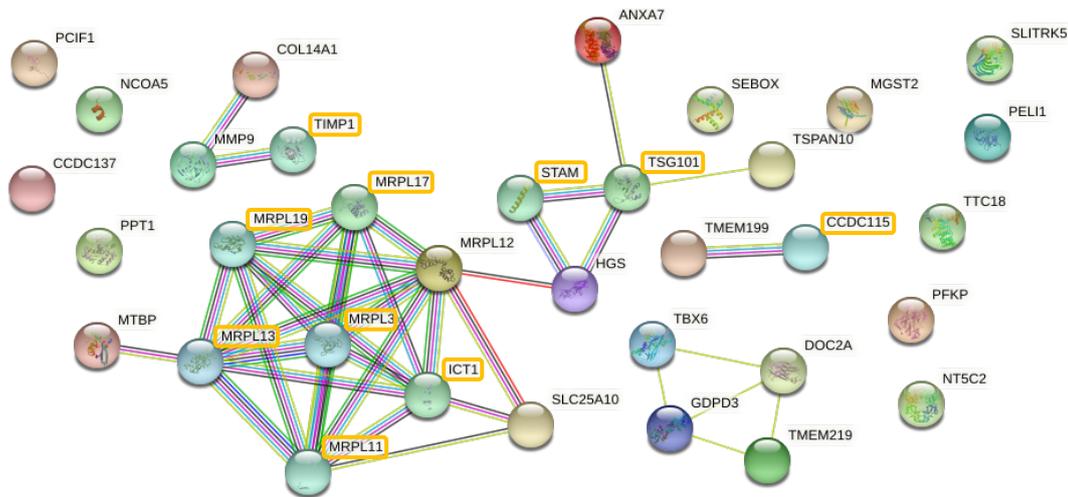
Fig. 2. Functional analysis of target genes. Protein-protein association network of the genes associated to active, genomically modified REMs from the STRING database [55]. The network was seeded with all genes in Table 1 and includes 10 additional interactors/proteins (marked with orange box) highly connected to the genes provided as input.

orectal carcinoma, enhances the resistance to doxorubicin [61]. Several reviews [62–64] point to the importance nuclear receptors play during the development of cancer, *e.g.* breast or prostate cancer, and their potential role as drug targets. We identified several nuclear receptors within our analysis, like the before mentioned NR2F2 and NR2C2 factors, but also RARA, NR3C1, NR2C1 and THRB.

A recent study [65] identified the G protein-coupled receptor 120 (GPR120) as a chemoresistance-promoting factor in breast cancer. The authors hypothesize that GPR120 mediates fatty acid synthesis and identified several factors, like SREBF1 and SREBF2, involved in this process. Due to the increased lipid synthesis this leads to a different lipid composition of the cell membrane and may prevent the drug from being absorbed.

Zheng *et al.* showed that the down regulation of RUNX3 in human lung adenocarcinoma is associated with docetaxel resistance by activating Akt1-mediated signaling [66]. Over-expression of RUNX3 leads to low AKT expression and higher treatment sensitivity.

ATF3 is a TF known to be involved in cellular stress response and is enriched in cells exposed to stress signals [67]. Under doxorubicin treatment, it has been reported that ATF3 affects cell death and cell cycle progression, however it is unclear whether the factor acts as a negative or positive regulator [68,69]. Nobori *et al.* claim that ATF3 plays a pivotal role as transcriptional regulator in the process of doxorubicin-induced cytotoxicity via an ERK-dependent pathway.

Walczynski *et al.* showed that the loss of ATF2 and ATF7, known to be homologous members in the AP-1 (activator protein 1) family, in MYC-expressing lymphoma cells in a mouse model are more resistant to doxorubicin induced apoptosis than control cells [70]. Further, they claim that ATF2/7 are important key TFs to suppress an oncogenic transformation. It has been found in other studies that upregulation of AP-1 expression confers resistance to chemotherapeutics [34].

8

rank	TF	FDR	rank	TF	FDR
1	**AP-1** [34]	3.83e-05	11	RARA	0.00062
2	**NR2C2** [60]	8.14e-05	12	NR3C1	0.00062
3	CREB3L4	0.00015	13	DBP	0.00079
4	HAND2	0.00016	14	**NR2F2** [61]	0.00084
5	**ATF3** [68, 69]	0.00017	15	NR2C1	0.00090
6	**ATF7** [70]	0.00018	16	GMEB2	0.001
7	MYBL1	0.00035	17	THRB	0.001
8	E2F8	0.00038	18	**SREBF2** [65]	0.0012
9	**ATF2** [70]	0.00039	19	**RUNX3** [66]	0.0013
10	NR2F1	0.00059	20	**SREBF1** [65]	0.0013

Table 2. Results of the TF enrichment analysis. TF binding in genomically modified active REMs was contrasted to random sets of active REMs in the same cells. Reported are the ranked TF names and FDR corrected enrichment p-values. TFs which have been reported to be linked to chemoresistance are in bold.

Discussion

We proposed a approach to identify genes whose expression might be affected by gRNAs targeting non-coding regulatory regions. To do so, we examined which gRNAs modify regulatory elements associated to potential target genes. The regulatory elements were retrieved from STITCHIT . Further, we took epigenetic data into account to identify REMs in open-chromatin or active enhancer regions of RPE-1 cells. Based on the target genes of active and genomically modified REMs, we constructed a protein-protein association network. Using the REM sequences, we performed a TF motif enrichment analysis to obtain key TFs involved in regulating doxorubicin resistance (see Fig. 1). Several of the identified target genes and TFs are already known to be associated to doxorubicin resistance, chemoresistance or cancer growth in general (marked in bold in Table 1 and Table 2).

Nevertheless, there are genes and TFs remaining that are worth to investigate more closely. For instance, the TF motif enrichment analysis points to the factor CREB3L4, for which we were not able to find an association to doxorubicin resistance. However, protein abundance of CREB3L1, which belongs to the same TF subfamily as CREB3L4, is predictive of the response of triple negative breast cancer to doxorubicin-based therapy [71]. In a recent paper, Pu *et al.* already identified that expression of CREB4L3 is increased in most breast cancer types, and that the downregulation leads to a decreasing proliferation, induces cell cycle arrest and apoptosis [72]. We observed a CREB3L4 binding site in a genomically modified REM associated to CCDC137. The gRNA (chr17:81,635,498) also affects MRPL12, SLC25A10, AC137896.1, HGS, TSPAN10 and AC139530.3.

We found one example, where a gRNA target site overlaps with the REM of the post-transcriptional regulator gene MIR5010 (Table 1). Mature microRNAs produced from this locus have the capacity

to regulate many target genes. While we were not able to find evidence that the mature microRNA products hsa-miR-5010-3p or hsa-miR-5010-5p of the MIR5010 gene are linked to chemoresistance against doxorubicin, the prevalence of hsa-miR-5010-3p may be involved in colon cancer recurrence. In a study of TNM-stage II colon cancer patients in two different cohorts, a four-miRNA recurrence prognosis signature was developed, including the abundance of hsa-miR-5010-3p [73]. Other than that, little is known about the regulatory roles of mature microRNAs of MIR5010. However, we queried the SpongeDB database, that holds predicted competing endogenous RNA interactions computed for different cancer types [74] using the Sponge algorithm [75]. We found that hsa-miR-5010-3p was predicted to regulate 3455 gene-gene ceRNA pairs in a pancancer model and many others in individual cancer types (Suppl. Tab. 1), suggesting an important regulatory role in cancer.

The overall association approach introduced here differs from other ad-hoc approaches. For example Wegner *et al.* identified target genes using the nearest gene approach, which links non-coding regions to adjacent genes without additional evidence. Our newly proposed approach is based on epigenetic data, first to infer the regulatory regions and second to detect genomically modified REMs, which are accessible in RPE-1 cells. Therefore, we conclude that our strategy is more informative than an approach, which is just based on genomic locations.

A CRISPR modification can create new TF binding sites in REMs. In rare cases, a consequence is that inactive REMs in the cell type of interest can become active by the gRNA caused modification. In general, no open chromatin data is available for the cell types after the modification. For that our current approach can not identify the REMs which become active by gaining a TF binding site.

We are aware that a CRISPR-Cas9 experiment can lead to off-target effects. Since we were able to link suggested target genes and TFs to chemoresistance events, we assume that at least for the considered cases here, off-target effects do not play a major role. Following the assumption that no strong off-target effects are causing chemoresistance in the experiment, it would be remarkable that only the modification of one REM leads to chemoresistance. This may fit to the observation that many of the active REMs detected here are likely regulating more than one target gene.

While we wanted to use ENCODE epigenomic data for RPE-1 cells, predictions of STITCHIT and additional open-chromatin and expression data for the Roadmap and Blueprint consortium can be directly obtained from our *EpiRegio* webserver [33]. *EpiRegio* holds data for diverse cell types, which allows to retrieve active REMs with their associated target genes.

To sum up, we introduced a cell-type specific approach to identify potential target genes of CRISPR modifications in non-coding regions, and useful downstream applications such as TF motif enrichment analysis. Several of our identified target genes and TFs are known to be associated to doxorubicin resistance, chemoresistance and cell growth in cancer cells. We believe that our approach based on

10

epigenetic data can be helpful to identify novel target genes for different CRISPR screens and various cell types.

Materials and methods

Analysis of Doxorubicin resistance in hTERT-RPE1 cells using CRISPR experiments

Here, we use a genome-wide CRISPR perturbation library consisting of partially randomized degenerated oligonucleotides (5'-NNDNNNNNHNNNNHDHNVVR-3') with flanking 3Cs homology regions which was created using ssDNA of template-plasmids and site-specific mutagenesis targeting coding and non-coding regions of the human genome in hTERT-RPE1 cells from ATCC (CRL-4000) [3]. Genomic coordiantes of the gRNAs have been obtained from this work (see Supp. Tab. 1).

Retrieval of epigenomic data

STITCHIT was applied to three data sets derived from different consortia, namely Roadmap [39], Blueprint and ENCODE [40]. We downloaded the REMs predicted for the 100kb window from ZENODO (DOI: 10.5281/zenodo.4316356).

Narrow peak files for human RPE-1 cells of a DNase1-seq and a H3K4me3 ChIP-seq data set were retrieved from ENCODE (accession numbers: ENCFF535GNR, ENCFF749ZBH, ENCFF922ERE). For the DNase1-seq the data contained 140,038 and 118,815 peaks for replicate 1 and 2, respectively. The replicates show a high agreement. The H3K4me3 ChIP-seq data included 34,179 peaks.

Identification of active and genomically modified REMs

First, we extended the gRNA target sites (see Sup. Table) to regions of length 200bp by adding 100bp up- and downstream. To identify the REMs which are active, and genomically modified by the extended gRNAs, we applied the following bedtools commands:

```
bedtools intersect -a <REMs> -b <gRNAs> -wa -u > <outputFile>
bedtools intersect -a <outputFile>  -b <H3K4me2-peaks> -wo
bedtools intersect -a <outputFile>  -b <DNase1-seq-rep1>, <DNase1-seq-rep1> -wo
```

Details about the motif enrichment analysis

The motif enrichment analysis is performed with *PASTAA* [58], which requires the DNA sequences of the 45 REMs and a set of known TF binding motifs. We determined the DNA sequences using the *getfasta* functionality in bedtools [76] and retrieved 515 TF binding motifs from the JASPAR motif database [77]. Additionally, *PASTAA* requests a ranking of the DNA sequences. Hence, we sorted the

sequences in descending order based on their maximal epigenetic signal from either the DNase1-seq or the H3K4me3 ChIP-seq, such that regions with higher read coverage occurred at the top. To ensure that PASTAA is able to discriminate enriched TF binding sites, we added 3 times as many randomly selected REMs (135) not affected by a gRNA, but with epigenomic evidence as being active in RPE1-cells. The background sequences are sorted in ascending order according to their epigenomic signal, and added to the bottom of the ranking the active, genomically modified REMs. We repeated the motif enrichment analysis 100 times, each run with a different background set, and averaged the TF enrichment results. The averaged result can be found in the Sup. Table.

Acknowledgements

Funding

References

1. M. Jinek, K. Chylinski, I. Fonfara *et al.*, "A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity," *Science*, vol. 337, no. 6096, pp. 816–821, 2012.

2. O. Shalem, N. E. Sanjana, E. Hartenian *et al.*, "Genome-scale crispr-cas9 knockout screening in human cells," *Science*, vol. 343, no. 6166, pp. 84–87, 2014.

3. M. Wegner, V. Diehl, V. Bittl *et al.*, "Circular synthesized CRISPR/Cas gRNAs for functional interrogations in the coding and noncoding genome," *Elife*, vol. 8, Mar 2019.

4. R. M. Meyers, J. G. Bryan, J. M. McFarland *et al.*, "Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells," *Nature genetics*, vol. 49, no. 12, pp. 1779–1784, 2017.

5. F. M. Behan, F. Iorio, G. Picco *et al.*, "Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens," *Nature*, vol. 568, no. 7753, pp. 511–516, 2019.

6. V. Pattanayak, S. Lin, J. P. Guilinger *et al.*, "High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity," *Nature Biotechnology*, vol. 31, no. 9, pp. 839–843, 2013.

7. J. G. Doench, N. Fusi, M. Sullender *et al.*, "Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9," *Nat. Biotechnol.*, vol. 34, no. 2, pp. 184–191, Feb 2016.

8. W. He, L. Zhang, O. D. Villarreal *et al.*, "De novo identification of essential protein domains from CRISPR-Cas9 tiling-sgRNA knockout screens," *Nature communications*, vol. 10, no. 1, pp. 1–10, 2019.

12

9. J. Y. Hsu, C. P. Fulco, M. A. Cole *et al.*, "Crispr-surf: discovering regulatory elements by deconvolution of crispr tiling screen data," *Nature methods*, vol. 15, no. 12, pp. 992–993, 2018.

10. M. A. Horlbeck, S. J. Liu, H. Y. Chang *et al.*, "Fitness effects of CRISPR/Cas9-targeting of long noncoding RNA genes," *Nature Biotechnology*, vol. 38, no. 5, pp. 573–576, 2020.

11. J. B. Wright and N. E. Sanjana, "CRISPR screens to discover functional noncoding elements," *Trends in Genetics*, vol. 32, no. 9, pp. 526–529, 2016.

12. L. Henkel, B. Rauscher, B. Schmitt *et al.*, "Genome-scale CRISPR screening at high sensitivity with an empirically designed sgRNA library," *BMC biology*, vol. 18, no. 1, pp. 1–21, 2020.

13. W. Li, H. Xu, T. Xiao *et al.*, "MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens," *Genome biology*, vol. 15, no. 12, p. 554, 2014.

14. W. Li, J. Köster, H. Xu *et al.*, "Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR," *Genome biology*, vol. 16, no. 1, p. 281, 2015.

15. B. Wang, M. Wang, W. Zhang *et al.*, "Integrative analysis of pooled CRISPR genetic screens using MAGeCK-Flute," *Nature protocols*, vol. 14, no. 3, pp. 756–780, 2019.

16. T. Hart and J. Moffat, "BAGEL: a computational framework for identifying essential genes from pooled library screens," *BMC bioinformatics*, vol. 17, no. 1, p. 164, 2016.

17. P. N. Spahn, T. Bath, R. J. Weiss *et al.*, "PinAPL-Py: a comprehensive web-application for the analysis of CRISPR/Cas9 screens," *Scientific reports*, vol. 7, no. 1, pp. 1–8, 2017.

18. S. Heinz, C. E. Romanoski, C. Benner, and C. K. Glass, "The selection and function of cell type-specific enhancers," *Nature Reviews Molecular Cell Biology*, vol. 16, no. 3, pp. 144–154, 2015.

19. Y. Zhang, C. H. Wong, R. Y. Birnbaum *et al.*, "Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations," *Nature*, vol. 504, no. 7479, pp. 306–310, 2013.

20. I. Krivega, R. K. Dale, and A. Dean, "Role of LDB1 in the transition from chromatin looping to transcription activation," *Genes and Development*, vol. 28, no. 12, pp. 1278–1290, 2014.

21. L. Yao, B. P. Berman, and P. J. Farnham, "Demystifying the secret mission of enhancers: Linking distal regulatory elements to target genes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 50, no. 6, pp. 550–573, 2015.

22. M. P. Creyghton, A. W. Cheng, G. G. Welstead *et al.*, "Histone H3K27ac separates active from poised enhancers and predicts developmental state," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 50, pp. 21 931–21 936, 2010.

23. H. H. He, C. A. Meyer, H. Shin *et al.*, "Nucleosome dynamics define transcriptional enhancers," *Nature Genetics*, vol. 42, no. 4, pp. 343–347, 2010.

24. G. Li, X. Ruan, R. K. Auerbach *et al.*, "Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation," *Cell*, vol. 148, no. 1-2, pp. 84–98, 2012.

25. C. D. Arnold, D. Gerlach, C. Stelzer *et al.*, "Genome-wide quantitative enhancer activity maps identified by STARR-seq," *Science*, vol. 339, no. 6123, pp. 1074–1077, 2013.

26. R. Andersson, C. Gebhard, I. Miguel-Escalada *et al.*, "An atlas of active enhancers across human cell types and tissues," *Nature*, vol. 507, no. 7493, pp. 455–461, 2014.

27. K. J. V. Nordström, F. Schmidt, N. Gasparoni *et al.*, "Unique and assay specific features of NOMe-, ATAC- and DNase I-seq data," *Nucleic Acids Research*, vol. 47, no. 20, pp. 10 580–10 596, 10 2019.

28. F. Schmidt, A. Marx, M. Hebel *et al.*, "Integrative analysis of epigenetics data identifies gene-specific regulatory elements," *bioRxiv*, 2019.

29. A. Visel, S. Minovitsky, I. Dubchak, and L. A. Pennacchio, "VISTA Enhancer Browser - A database of tissue-specific human enhancers," *Nucleic Acids Research*, vol. 35, no. SUPPL. 1, pp. 88–92, 2007.

30. J. Wang, X. Dai, L. D. Berry *et al.*, "HACER: An atlas of human active enhancers to interpret regulatory variants," *Nucleic Acids Research*, vol. 47, no. D1, pp. D106–D112, 2019.

31. S. Fishilevich, R. Nudel, N. Rappaport *et al.*, "GeneHancer: genome-wide integration of enhancers and target genes in GeneCards," *Database : the journal of biological databases and curation*, vol. 2017, pp. 1–17, 2017.

32. Z. Cai, Y. Cui, Z. Tan *et al.*, "RAEdb: A database of enhancers identified by high-throughput reporter assays," *Database*, vol. 2019, no. 4, pp. 1–5, 2019.

33. N. Baumgarten, D. Hecker, S. Karunanithi *et al.*, "EpiRegio: analysis and retrieval of regulatory elements linked to genes," *Nucleic Acids Research*, vol. 48, no. W1, pp. W193–W199, 05 2020.

34. Z. Wang, Q. Zhang, W. Zhang *et al.*, "HEDD: Human Enhancer Disease Database," *Nucleic Acids Research*, vol. 46, no. D1, pp. D113–D120, 2018.

35. G. Zhang, J. Shi, S. Zhu *et al.*, "DiseaseEnhancer: A resource of human disease-associated enhancer catalog," *Nucleic Acids Research*, vol. 46, no. D1, pp. D78–D84, 2018.

36. K. M. Tewey, T. C. Rowe, L. Yang *et al.*, "Adriamycin-induced DNA damage mediated by mammalian DNA topoisomerase II," *Science*, vol. 226, no. 4673, pp. 466–468, Oct 1984.

37. S. Cagnol and J. C. Chambard, "ERK and cell death: mechanisms of ERK-induced cell death–apoptosis, autophagy and senescence," *FEBS J.*, vol. 277, no. 1, pp. 2–21, Jan 2010.

38. H. G. Stunnenberg, S. Abrignani, D. Adams *et al.*, "The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery," *Cell*, vol. 167, no. 5, pp. 1145–1149, 2016.

39. Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman *et al.*, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–329, 2015.

40. T. E. P. Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012.

41. C. L. Dieck and A. Ferrando, "Genetics and mechanisms of NT5C2-driven chemotherapy resistance in relapsed ALL," *Blood*, vol. 133, no. 21, pp. 2263–2268, may 2019.

42. J. Shen, Z. Jin, H. Lv *et al.*, "PFKP is highly expressed in lung cancer and regulates glucose metabolism," *Cellular Oncology*, vol. 43, no. 4, pp. 617–629, 2020.

43. H. Liu, D. Guo, Y. Sha *et al.*, "ANXA7 promotes the cell cycle, proliferation and cell adhesion-mediated drug resistance of multiple myeloma cells by up-regulating CDC5L," *Aging*, vol. 12, no. 11, pp. 11 100–11 115, 2020.

44. Y. Gao, B. Sun, J. Hu *et al.*, "Identification of gene modules associated with survival of diffuse large B-cell lymphoma treated with CHOP-based chemotherapy," *The Pharmacogenomics Journal*, vol. 20, no. 5, pp. 705–716, 2020.

14

45. Q. Cai, M. Dozmorov, and Y. Oh, "IGFBP-3/IGFBP-3 Receptor System as an Anti-Tumor and Anti-Metastatic Signaling in Cancer," *Cells*, vol. 9, no. 5, p. 1261, May 2020.

46. M. D. Indermaur, Y. Xiong, S. G. Kamath *et al.*, "Genomic-directed targeted therapy increases endometrial cancer cell sensitivity to doxorubicin," *American Journal of Obstetrics and Gynecology*, vol. 203, no. 2, pp. 158.e1 – 158.e40, 2010.

47. A. S. Baras, N. Gandhi, E. Munari *et al.*, "Identification and validation of protein biomarkers of response to neoadjuvant platinum chemotherapy in muscle invasive urothelial carcinoma," *PLOS ONE*, vol. 10, no. 7, pp. 1–11, 07 2015.

48. A. Buckley, B. Bibby, M. Dunne *et al.*, "Characterisation of an Isogenic Model of Cisplatin Resistance in Oesophageal Adenocarcinoma Cells," *Pharmaceuticals*, vol. 12, no. 1, p. 33, Feb 2019.

49. L. Rochette, A. Meloux, M. Zeller *et al.*, "Mitochondrial SLC25 Carriers: Novel Targets for Cancer Therapy," *Molecules*, vol. 25, no. 10, p. 2417, May 2020.

50. G. Wang, J. Xia, C. Chen *et al.*, "SLC25A10 performs an oncogenic role in human osteosarcoma," *Oncol Lett*, vol. 20, no. 4, p. 2, 2020.

51. P. Spallarossa, P. Altieri, S. Garibaldi *et al.*, "Matrix metalloproteinase-2 and -9 are induced differently by doxorubicin in H9c2 cells: The role of MAP kinases and NAD(P)H oxidase," *Cardiovasc. Res.*, vol. 69, no. 3, pp. 736–745, Feb 2006.

52. E. Dvash, M. Har-Tal, S. Barak *et al.*, "Leukotriene C4 is the major trigger of stress-induced oxidative DNA damage," *Nature Communications*, vol. 6, no. 1, Dec. 2015.

53. K. Sun, S. Wang, J. He *et al.*, "NCOA5 promotes proliferation, migration and invasion of colorectal cancer cells via activation of PI3k/AKT pathway," *Oncotarget*, vol. 8, no. 64, pp. 107 932–107 946, Nov. 2017.

54. J.-S. Moon, H. E. Kim, E. Koh *et al.*, "Krüppel-like factor 4 (KLF4) activates the transcription of the gene for the platelet isoform of phosphofructokinase (PFKP) in breast cancer," *The Journal of biological chemistry*, vol. 286, no. 27, pp. 23 808–23 816, jul 2011.

55. D. Szklarczyk, A. L. Gable, K. C. Nastou *et al.*, "The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets," *Nucleic Acids Research*, 11 2020, gkaa1074.

56. T. Arbogast, A.-M. Ouagazzal, C. Chevalier *et al.*, "Reciprocal effects on neurocognitive and metabolic phenotypes in mouse models of 16p11.2 deletion and duplication syndromes," *PLOS Genetics*, vol. 12, no. 2, pp. 1–35, 02 2016.

57. X. Zheng, F. Y. Demirci, M. M. Barmada *et al.*, "A Rare Duplication on Chromosome 16p11.2 Is Identified in Patients with Psychosis in Alzheimer's Disease," *PLOS ONE*, vol. 9, no. 11, pp. 1–7, 11 2014.

58. H. G. Roider, T. Manke, S. O'Keeffe *et al.*, "PASTAA: identifying transcription factors associated with sets of co-regulated genes," *Bioinformatics*, vol. 25, no. 4, pp. 435–442, 12 2008.

59. X. Ding, D.-R. Yang, L. Xia *et al.*, "Targeting TR4 nuclear receptor suppresses prostate cancer invasion via reduction of infiltrating macrophages with alteration of the TIMP-1/MMP2/MMP9 signals," *Molecular Cancer*, vol. 14, no. 1, p. 16, 2015.

60. L. Hu, Y. Sun, J. Luo *et al.*, "Targeting TR4 nuclear receptor with antagonist bexarotene increases docetaxel sensitivity to better suppress the metastatic castration-resistant prostate cancer progression," *Oncogene*, vol. 39, no. 9, pp. 1891–1903, 2020.

61. X. Wang, R. Jiang, W. Feng *et al.*, "COUP-TFII suppresses colorectal carcinoma resistance to doxorubicin involving inhibition of epithelial-mesenchymal transition," *American journal of translational research*, vol. 8, no. 9, pp. 3921—-3929, Sep. 2016.

62. L. Zhao, S. Zhou, and J.-Å. Gustafsson, "Nuclear receptors: recent drug discovery for cancer therapies," *Endocrine Reviews*, Mar. 2019.

63. R. B. Riggins, M. M. Mazzotta, O. Z. Maniya, and R. Clarke, "Orphan nuclear receptors in breast cancer pathogenesis and therapeutic response," *Endocrine-Related Cancer*, vol. 17, no. 3, pp. R213–R231, Sep. 2010.

64. D. Wu, A. Cheung, Y. Wang *et al.*, "The emerging roles of orphan nuclear receptors in prostate cancer," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1866, no. 1, pp. 23–36, Aug. 2016.

65. X. Wang, S. He, Y. Gu *et al.*, "Fatty acid receptor GPR120 promotes breast cancer chemoresistance by upregulating ABC transporters expression and fatty acid synthesis," *EBioMedicine*, vol. 40, pp. 251–262, Feb. 2019.

66. Y. Zheng, R. Wang, H.-Z. Song *et al.*, "Epigenetic downregulation of RUNX3 by DNA methylation induces docetaxel chemoresistance in human lung adenocarcinoma cells by activation of the AKT pathway," *The International Journal of Biochemistry Cell Biology*, vol. 45, no. 11, pp. 2369 – 2378, 2013.

67. T. Hai, C. D. Wolfgang, D. K. Marsee *et al.*, "ATF3 and stress responses," *Gene Expr.*, vol. 7, no. 4-6, pp. 321–335, 1999.

68. K. Nobori, H. Ito, M. Tamamori-Adachi *et al.*, "ATF3 inhibits doxorubicin-induced apoptosis in cardiac myocytes: a novel cardioprotective role of ATF3," *J. Mol. Cell. Cardiol.*, vol. 34, no. 10, pp. 1387–1397, Oct 2002.

69. E. J. Park, H. K. Kwon, Y. M. Choi *et al.*, "Doxorubicin induces cytotoxicity through upregulation of pERK-dependent ATF3," *PLoS ONE*, vol. 7, no. 9, p. e44990, 2012.

70. J. Walczynski, S. Lyons, N. Jones, and W. Breitwieser, "Sensitisation of c-MYC-induced b-lymphoma cells to apoptosis by ATF2," *Oncogene*, vol. 33, no. 8, pp. 1027–1036, Feb. 2013.

71. B. Denard, S. Jiang, Y. Peng, and J. Ye, "CREB3l1 as a potential biomarker predicting response of triple negative breast cancer to doxorubicin-based chemotherapy," *BMC Cancer*, vol. 18, no. 1, Aug. 2018.

72. Q. Pu, L. Lu, K. Dong *et al.*, "The novel transcription factor CREB3l4 contributes to the progression of human breast carcinoma," *Journal of Mammary Gland Biology and Neoplasia*, vol. 25, no. 1, pp. 37–50, Feb. 2020.

73. H. Jacob, L. Stanisavljevic, K. E. Storli *et al.*, "A four-microRNA classifier as a novel prognostic marker for tumor recurrence in stage II colon cancer," *Scientific Reports*, vol. 8, no. 1, p. 6157, 2018.

74. M. Hoffmann, E. Pachl, M. Hartung *et al.*, "SPONGEdb: a pan-cancer resource for competing endogenous RNA interactions," *NAR Cancer*, vol. 1, Jan. 2020.

75. M. List, A. Dehghani Amirabad, D. Kostka, and M. H. Schulz, "Large-scale inference of competing endogenous RNA networks with sparse partial correlation," *Bioinformatics*, vol. 35, no. 14, pp. i596–i604, jul 2019.

16

76. A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar 2010.

77. A. Khan, O. Fornes, A. Stigliani *et al.*, "JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework," *Nucleic Acids Research*, vol. 46, no. D1, pp. D260–D266, Nov. 2017.