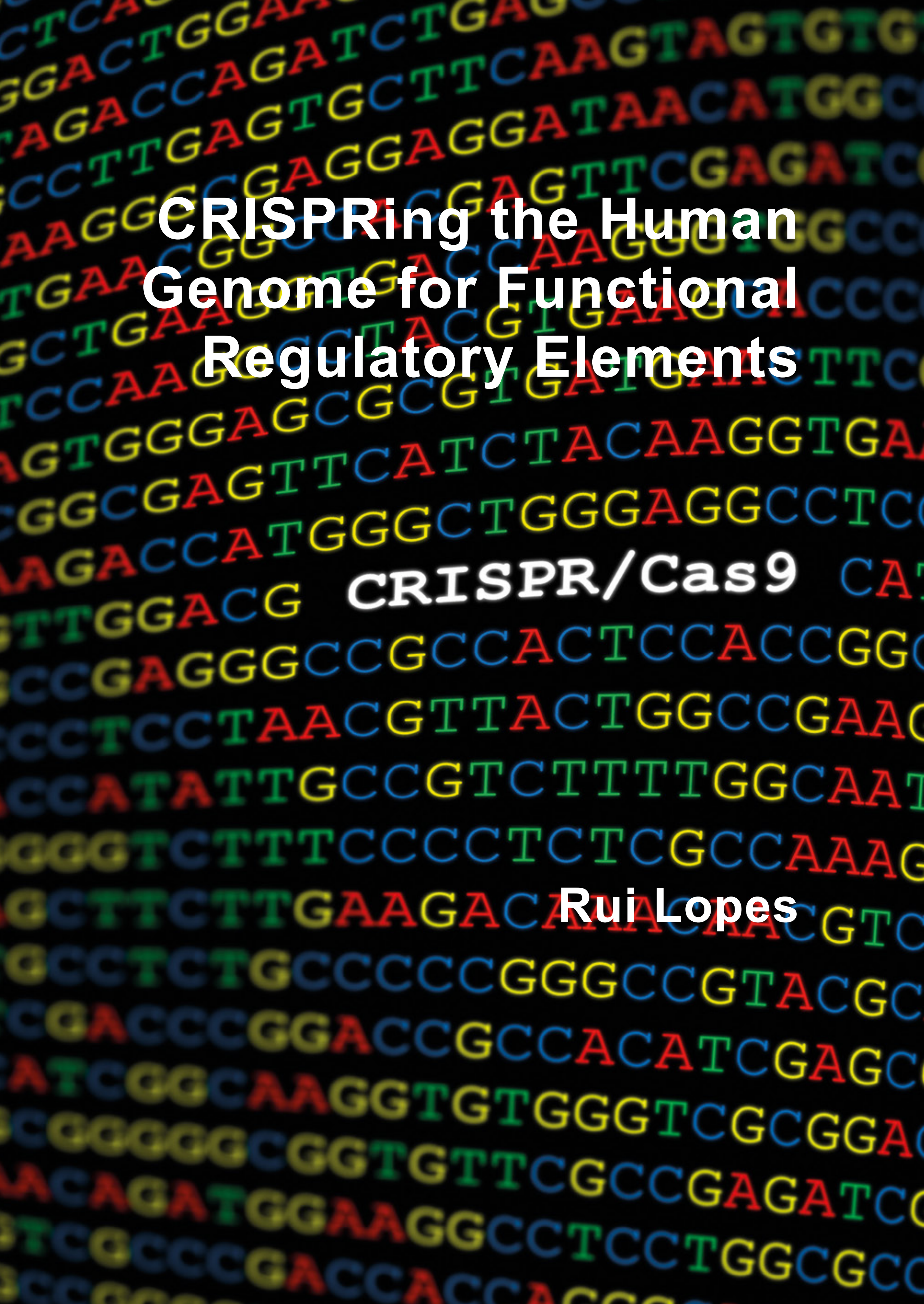


CRISPRing the Human Genome for Functional Regulatory Elements

Rui Lopes

"*The known is finite, the unknown infinite; intellectually we stand on an islet in the midst of an illimitable ocean of inexplicability. Our business in every generation is to reclaim a little more land, to add something to the extent and the solidity of our possessions*" T.H. Huxley

ISBN/EAN: 978-94-028-0917-6

Copyright of the cover image - Catalinr ID 97763119, Dreamstime

CRISPRing the Human Genome for Functional Regulatory Elements

Het verCRISPRen van het menselijk genoom voor functionele regulerende elementen

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by the command of the
Rector Magnificus

Prof.dr. H.A.P. Pols

and in accordance with the decision of the Doctorate Board.
The public defense shall be held on

Friday 9th of March 2018 at 11h30m

by

Rui Filipe Marques Lopes
born in Mangualde da Serra,
Portugal

Erasmus University Rotterdam

Doctoral committee:

Supervisor: Prof.dr. R. Agami

Other members: Prof.dr. J.H. Gribnau
Prof.dr. H.R. Delwel
Dr. R. Elkon

Co-supervisor: Dr. G. Korkmaz

Table of Contents

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Appendix

Abbreviations

bp	Base pair
Cas9	CRISPR associated protein 9
ChIA-PET	Chromatin interaction analysis with paired-end tag sequencing
ChIP-seq	Chromatin immunoprecipitation sequencing
ChIRP-seq	Chromatin isolation by RNA purification sequencing
CRISPR	Clustered regularly interspaced short palindromic repeats
DSBs	Double-strand breaks
dCas9	Catalytically-inactive Cas9
DNA	Deoxyribonucleic acid
DNase I	Deoxyribonuclease I
DNase-seq	DNase hypersensitivity sequencing
eRNA	Enhancer-associated RNA
ESC	Embryonic stem cell
FISH	Fluorescence *in situ* hybridization
GCR	Global control region
GRO-seq	Global run-on sequencing
GWAS	Genome-wide association studies
HDR	Homology-directed repair
Indels	Insertions and deletions
kb	Kilo base pair
LCR	Locus control region
lncRNA	Long non-coding RNA
Mb	Mega base pair
MPRA	Massive parallel reporter assay
NGS	Next-generation sequencing
NHEJ	Non-homologous end joining
PAM	Protospacer adjacent motif
RNA	Ribonucleic acid
RNAi	RNA interference
RNAPII	RNA polymerase II
RNA-seq	RNA sequencing
sgRNA	Single guide RNA
SNP	Single-nucleotide polymorphism
STARR-seq	Self-transcribing active regulatory region sequencing
SV40	Simian virus 40
TAD	Topological associating domain
TALE	Transcription activator-like effector
TF	Transcription factor
ZFN	Zinc finger nuclease

Chapter 1

General introduction

Transcriptional regulation

The genomic DNA sequence carries information in two fundamental forms: first, in transcribed genes that specify mRNAs and other functional non-coding RNAs; second, in regulatory sequences that control the expression levels and patterns of those genes. The initial paradigms of gene regulation were established by studying transcription in prokaryotes, which mainly rely on promoter-proximal DNA sequences to control transcription[1]. In unicellular organisms, this information can determine absolute levels of transcription and also mediate gene expression changes in response to external stimuli. Metazoan organisms present a challenge in this regard since a single cell originates diverse cell-types, which have distinct morphology and function, and constitute the different structures present in an adult multicellular organism. The advent of next-generation sequencing (NGS) technologies enabled comparing the genomes of different species, and these studies revealed that organismal complexity and genome size do not correlate in a linear manner[2]. Therefore, morphological and developmental complexity are not a direct product of increased number of genes but, instead, of alternative mechanisms. Notably, complexity can arise by diversifying the patterns of gene expression, both in space and time, within an organism. In metazoans, transcriptional control is dependent not only on promoters but also on distal *cis*-regulatory elements known as enhancers. The uncoupling of enhancers from their target promoter was first demonstrated when Banerji *et al*. showed that the SV40 enhancer is able to increase the expression of a heterologous gene (*β-globin*) over a distance of 10 kb[3]. Recent studies provided dramatic examples of very long-range interactions between enhancers and promoters in vertebrate genomes. For example, the expression of *SHH* and *MYC* is regulated by distal enhancers that map more than 1 Mb from their promoter region[4,5]. The regulation of promoters by enhancers at a distance opens the door for complex transcriptional regulation, whereby a gene can be differentially expressed in distinct cell-types and in response to different environmental cues. A well-studied example is the regulation of *even-skipped* in *Drosophila*, which is expressed in seven distinct stripes along the length of the embryo due to the action of five different enhancers[6]. Thus, it is very likely that the distal location and modular organization of enhancers enabled the development of multiple cell-types and contributed to the evolutionary diversity of metazoans.

Hallmarks of enhancer elements

Enhancers were first characterized by gain-of-function reporter assays in immortalized cell lines[3,7]. These seminal studies defined enhancers as DNA sequences that can activate transcription independently of their distance and orientation relative to the target promoter. This flexibility is a hallmark of enhancer elements and remains part of their functional definition to date. Enhancers are commonly located in intergenic regions or within the introns of protein-coding genes. However, this flexibility poses a great challenge to catalog the full set of enhancers present in the human genome. Whereas promoters can be identified simply by sequencing the 5' end of genes, no such clear-cut criterion exists that can locate an enhancer and its target gene(s).

A central feature of enhancers is their ability to function as binding platforms for transcription factors (TFs). The DNA sequence of enhancers is usually 200-500 bp long and contains clustered recognition sites for multiple TFs. The conservation of these sequences is often used to identify putative enhancers[8], and several studies indicate that their activity is largely cell-type and specie specific[9]. In general, several TFs are required for the activation of enhancers, including lineage-specific and signal-responsive factors that ensure the integration of intrinsic and extrinsic cues at these elements. The ability of TFs to activate transcription on chromatin templates is dependent on the recruitment of coactivator proteins, such as p300[10,11]. These factors often lack DNA-binding capacity, but instead function as histone modifiers, chromatin remodelers or recruiters of general TFs and RNAPII. Surprisingly, it was found that general TFs and RNAPII also bind to enhancer regions, leading to the production of enhancer-associated RNAs (eRNAs)[12,13]. The expression of eRNAs correlates with enhancer activity and there is abundant evidence supporting a role for these transcripts in gene regulation (see section

“Functional roles of eRNAs”). The binding of TFs at enhancers is associated with regions depleted of nucleosomes that are highly sensitive to DNA nucleases like DNase I[14]. However, nucleosomes immediately adjacent to enhancer regions are marked with specific histone modifications, namely H3K4me1 and H3K27Ac[15-17]. Notably, H3K4me3 is associated with gene promoters, which usually exhibit low levels of H3K4me1 at the transcription start site. These “chromatin signatures”, often in combination with DNase I hypersensitivity and coactivator binding, are frequently used to annotate enhancers in a genome-wide scale[18-21]. Based on such experiments, it was suggested that there are approximately one million enhancers in the human genome[20,21]. However, dozens of histone modifications remain to be tested and, therefore, a comprehensive census of enhancers based on chromatin signatures remains a subject of speculation.

Mechanisms of enhancer function

Enhancers play a central role in controlling spatiotemporal gene expression, which is essential to specify different cell lineages during development (reviewed in[6]). Still, the nature of enhancer-promoter communication is one of the outstanding mysteries of transcriptional regulation. More than 30 years passed since the discovery of the archetype SV40 enhancer, and yet, we do not fully understand the mechanisms of this process. It is generally accepted that enhancers activate transcription by delivering essential factors to the gene promoter, which stimulate the formation of the preinitiation complex (PIC) or the transition from initiation to elongation. There are several models that try to explain how enhancers communicate with promoters over long distances[22], but two of them stand out from the remaining: “looping” and “tracking”. The first postulates that enhancers and promoters interact directly while the intervening DNA sequence is looped out[23]. The latter proposes that enhancers diffuse along the chromatin fiber in search of a target promoter[24]. Nevertheless, both models agree that the mechanism of action of enhancers requires direct interactions with the gene promoter. In recent years, the looping model as received abundant support through the results obtained by Chromosome Conformation Capture (3C) and its derivatives (4C, 5C and Hi-C)[25-28]. These studies revealed that enhancers and promoters are extensively engaged in interactions within multiple loci in mammalian genomes[29]. The fact that enhancers often colocalize with the promoters they regulate was interpreted as the result of direct enhancer-promoter interactions, which are required for the activation of gene expression. This hypothesis is supported by several studies that found a strong correlation between active transcription and enhancer-promoter interactions. For example, knockout of TFs that are required for *β-globin* expression results in the loss of colocalization of the gene promoter with its locus control region (LCR)[30,31]. Additionally, it was shown that some enhancers can exhibit a preference for specific classes of promoters, such as the ones containing a canonical TATA box[32,33], further supporting a direct communication between these regulatory elements. Nonetheless, it is not clear whether the spatial colocalization of enhancer and promoter regions is a cause or consequence of gene expression. Deng *et al*. addressed this question by tethering Ldb1 to the promoter of *β-globin* via an artificial zinc finger (ZF) protein[34,35]. They found that ZF-Ldb1 was sufficient to establish a loop between *β-globin* and its LCR, recruit RNAPII and activate gene expression. These results support a causal role for DNA looping in gene activation and demonstrate that forced chromatin interactions can overcome tightly regulated developmental mechanisms.

Topology of enhancers and their regulatory landscapes

Evidence supporting enhancer-promoter interactions are part of a bigger picture showing that nuclear organization is a major determinant of gene expression. Imaging experiments revealed that interphase chromosomes tend to occupy discrete areas, called “chromosome territories”, rather than spreading throughout the nucleus[36]. Furthermore, individual chromosomes are organized in series of topologically associating domains (TADs), which are megabase-sized regions containing 5-10 genes and a few hundred enhancers[37,38]. TADs have similar boundaries in all human cell-types examined to date and display high frequency of self-interactions as measured by Hi-C[39-41]. Accordingly, TADs have been proposed to constrain enhancer-promoter interactions because the vast majority of DNA contacts occur within the TADs[39,41]. This hypothesis explains why enhancer-promoter interactions mainly occur in *cis* and are limited in length within a chromosome. Still, it does not answer the question: how do enhancers communicate with the right promoter(s) in time and space?

In the nucleus, gene loci can colocalize on the basis of shared associations with specific factors, such as RNAPII. Visualization of RNAPII or nascent mRNAs suggested that transcription is localized to a limited number of foci, known as “transcription factories”[42,43]. This term was proposed to explain the observation that active gene loci located in the same or even separate chromosomes tend to colocalize in the nucleus. In addition to RNAPII, other factors are organized into discrete foci and can, either directly or indirectly, bring distal loci into close proximity with each other. In this regard, CTCF (CCCTC-Binding Factor) has emerged as a key player in chromatin organization and gene regulation. CTCF is a transcriptional regulator that binds DNA through its ZF domains. Strikingly, it is the only known protein to bind to insulators (also known as boundary elements) and mediate this type of activity in vertebrates[44]. The main function of insulators is to block genes from being affected by the transcriptional activity of neighboring loci. Therefore, they limit the action of transcriptional regulatory elements to defined regions, and effectively partition the genome into discrete realms of expression. The activity of insulators is mainly defined by their ability to block enhancer-promoter communication and prevent spreading of heterochromatin (reviewed in[45]). CTCF can also associate with itself[46], and these CTCF-CTCF interactions have been implicated in the formation of chromatin loops as detected by 3C-based techniques[47,48]. Interestingly, CTCF associates with cohesin and this seems to be required for insulating activity[49,50]. A study by Kagey *et al*. found that enhancers and promoters are associated with cohesin and mediator[51], providing a potential mechanistic link between long-range CTCF-CTCF interactions and enhancer-promoter communication. In recent years, this hypothesis has gained momentum due to the availability of genome-wide maps of the proteins that bind enhancers, promoters and insulators, together with information about the physical interactions that occur between them[52-55]. This information gave rise to a model in which each chromosome contains thousands of DNA loops, formed by the interaction of two CTCF molecules bound to different loci and reinforced by a cohesin ring. The proteins that bind to enhancers within the loop are constrained such that they tend to interact only with promoters in their vicinity. These CTCF-CTCF loops have been termed “insulated neighborhoods” because they insulate enhancers and genes within the loop from enhancers and genes outside the loop (reviewed in[56]).

Several lines of evidence support a function for insulated neighborhoods in activation and repression of gene expression. First, the majority of enhancer-promoter interaction occur within insulated neighborhoods (e.g. ~90% in human ESCs)[54,55,57,58]. Second, genetic or epigenetic perturbation of neighborhood boundaries leads to changes in local gene expression[55,57-60]. Finally, somatic mutations in CTCF-binding sequences overlapping with neighborhood boundaries were found in multiple tumor-types[57,59,61]. The insulated neighborhood model suggests an explanation for how enhancer-promoter specificity is obtained when a single gene occurs together with its regulatory elements within the neighborhood. However, it does not fully justify enhancer-promoter specificity when there are multiple genes within the loop. It was estimated that in neighborhoods with two genes, their expression patterns are concordant in ~60% of the cases (i.e. both are active or both are silent), suggesting that they are co-regulated[56]. In *Drosophila*, there is evidence that an enhancer can target all genes within an insulated neighborhood[62]. Nonetheless, it is very likely that enhancer-promoter communication is determined, to a great extent, by the interaction of specific factors bound at these elements[30-35].

Functional roles of eRNAs

Several reports over the past half a century hinted at the existence of short-lived RNA species in the nucleus. In 1959, it was found that the majority of nascent RNA is rapidly degraded and does not contribute to the pool of mRNAs[63]. However, it was only in the 1990’s that specific transcription at enhancers was documented in the LCR of the globin genes[64-66]. Additional examples were found later at the LCRs of *MHC Class II*[67] and *GH1*[68]. However, widespread RNA transcription at enhancers only became apparent in recent years through the application of NGS technologies. Using total RNA-seq, Kim *et al*. found a broad pattern of transcription at active enhancers in neuronal cells[12]. Moreover, several studies identified RNAPII complexes enriched at putative enhancer regions by using ChIP-seq[12,13,69]. The discovery of pervasive transcription at enhancers indicates that eRNAs, in addition to introns, are important contributors to the lowly-stable pool of nuclear RNA[70]. Genome-wide detection of nascent RNA by Global Run-on sequencing (GRO-seq) demonstrated that eRNAs are widely expressed in macrophages, breast, colon and prostate cancer cells[71-75]. Several studies reported that expression of eRNAs is responsive to extrinsic cues[12,71,74,76], suggesting a role for these molecules in the regulation of gene expression. Indeed, it was demonstrated that transcription of eRNAs preceded the activation of target genes[13,77] and their expression correlated with the expression of neighboring genes[12,13]. Additional evidence showed that eRNA expression is specifically regulated by signal-dependent TFs, such as p53 and ERα, and is highly correlated with changes in expression of target genes[74,78]. A causal role for eRNAs in transcriptional activation was demonstrated in a number of subsequent studies, in which the depletion of eRNAs led to specific repression of target genes in human cells[74,78-81]. It was also shown that transcriptional activation could be recapitulated in reporter assays, and this was dependent on the expression of eRNAs[74,78,79]. Moreover, Lam *et al.* provided evidence that reporter vectors containing eRNA-coding sequences have higher transcriptional activity compared to the ones containing the enhancer sequence alone[81]. The eRNA sequence seems to be important *per se* since the increased expression was abolished upon reversing its orientation relative to the enhancer[81]. Collectively, these studies indicate that expression of eRNAs is a hallmark of active enhancer elements and support a main role for them in transcriptional regulation.

Mechanisms of eRNA function

eRNAs were initially defined as non-coding RNAs produced from putative enhancer regions marked by high H3K4me1, low H3K4me3, and occupied by RNAPII[12,13]. Still, they are a poorly defined class of RNAs that is associated with different features and mechanisms of action. In general, eRNAs have a 5' cap but are not spliced or polyadenylated[70,81]. Polyadenylated eRNAs are usually transcribed as a unidirectional unit, although enhancers with bidirectional transcription and non-polyadenylated transcripts are more common[69]. The half-life of eRNAs is low compared to mRNAs and long non-coding RNAs (lncRNAs), but their transcription initiation frequency is similar to that of protein-coding genes[81]. These features suggested that eRNAs have a nuclear function, and several mechanisms have been proposed to explain how eRNAs might contribute to gene regulation. It was observed that transcription activity at the *β-globin* LCR correlated with its sensitivity to DNase I, hinting that intergenic non-coding RNAs can play a role in the maintenance of active chromatin states[82]. Additionally, Mousavi *et al*. proposed that eRNAs facilitate RNAPII recruitment the target promoter(s)[83]. They showed that eRNAs were critical for the expression of *MyoD*, and that their knockdown decreased RNAPII occupancy at the promoter but not at the enhancer. This is in agreement with earlier observations at the HS2 enhancer: inhibition of RNAPII elongation results in decreased recruitment of RNAPII to the *β-globin* promoter but not at the HS2 enhancer[84]. This evidence also suggests that recruitment of RNAPII and transcription at enhancers is an early event and precedes the activation of target genes. Genome-wide studies of chromatin interactions revealed that enhancers engaged in looping with promoters express higher levels of eRNAs[85,86]. Moreover, eRNAs interact both with mediator[80] and cohesin complexes[74], suggesting that they might be involved in the establishment or maintenance of chromosome conformation. Importantly, depletion of eRNAs caused a strong decrease in enhancer-promoter interactions and a concomitant reduction of target gene expression[74,80]. Available data indicates that this might not be a general mechanism since enhancer-promoter interactions do not always require eRNAs (see General discussion). eRNAs might also exert their function by acting as a decoy for the negative elongation factor (NELF) complex. It was demonstrated that eRNAs are synthesized prior to target gene transcription and interact with a subunit of NELF[87]. Knockdown of eRNAs impaired the release of NELF from target promoters, which coincided with downregulation of gene expression[87]. Together, these studies demonstrated that eRNAs are involved in almost all stages of transcriptional activation, from chromatin accessibility and loop formation to RNAPII loading and pause release. Future studies should aim at identifying the protein partners of eRNAs in order to provide comprehensive insights into their mechanisms of action.

The role of enhancers in disease

Enhancers are essential for orchestrating complex gene expression patterns that are required for the proper development of adult organisms. Therefore, it is not surprising that dysfunction of enhancers or the factors that bind to them is an important component in human disease. As mentioned above, enhancers translate extracellular signals to an intracellular response in the form of changes in gene expression. In general, this happens through a cascade of signaling events that culminate in the nucleus through the action TFs. A large number of cancer-associated genes are TFs or kinases that control their activity, and therefore, it is not surprising that many gene regulatory circuits are altered in cancer[88]. One of the most striking examples is p53, which is a TF that activates gene expression in response to diverse cellular stresses, thereby leading to DNA repair, cell cycle arrest and apoptosis (reviewed in[89]). Not surprisingly, *p53* is the most frequently altered gene in human tumors, with mutation rates ranging from ~10% up to nearly 100%[90]. The vast majority of mutations are located in the DNA binding domain of p53 - thereby impairing its functions as a TF and tumor-suppressor. Interestingly, genome-wide studies showed that a large fraction of p53 binding sites overlap with distal regulatory elements[75,76,78], suggesting that p53 regulates its target genes by binding to enhancers. Another example is ERα, which is a ligand-dependent transcription factor that promotes cell growth. ERα is activated by estradiol, which is its natural ligand, or through phosphorylation events mediated by kinases such as MAPK/PI3K[91]. ERα is expressed in ~70% of breast tumors and, therefore, it is a major target for hormonal therapy in this type of cancer[91]. Genome-wide analysis of ERα binding by ChIP-seq identified many events at intergenic and intronic regions that display typical features of enhancers[92,93]. The vast majority of tumors that relapse after hormonal therapy still express ERα[94], underlining the importance of identifying the enhancers and target genes of this pathway. In recent years, a number of inhibitors were developed to target transcriptional regulators that bind enhancers. In particular, the use BET inhibitors for cancer treatment has generated great enthusiasm and their effect is currently under evaluation in clinical trials[95]. TFs are considered the *Holy Grail* of cancer therapy and, for many years, it was thought that they were undruggable. Their remarkable diversity and potency as drivers of tumorigenesis justifies a continued pursuit of novel drugs to target TFs.

Similar to mutations in protein-coding genes, variation in enhancer sequences has been causally associated with several monogenic disorders (reviewed in[96]). A notable example is the dysregulation of *SHH* expression and limb malformations. The expression *SHH* is governed by a distal enhancer element, known as the ZPA regulatory sequence (ZRS), located approximately 1 Mb away from its promoter. Point mutations within the ZRS have been linked to a congenital disease characterized by the formation of extra digits[97], whereas deletion of the entire element causes truncation of limbs in mice[98]. Additional examples include mutations in the enhancers of *Sox9* and *Tbx5* that cause Pierre Robin anomaly[99] and congenital heart disease[100], respectively.

The main evidence connecting genetic alterations in enhancers and cancer comes from GWAS. To date, these studies have identified more than 400 SNPs that significantly predispose individuals to various types of cancer[101]. Interestingly, the vast majority of disease-associated variants map to non-coding regions of the genome: 40% are intergenic and a similar percentage map to intronic regions[102,103]. A large fraction of cancer-risk SNPs occur in regions enriched in expression quantitative trait loci (eQTLs)[104], DNase I hypersensitive sites[105] and eRNA expression[106], which are features indicative of enhancers. In recent years, a number of studies showed that genetic variation at enhancers can predispose individuals to cancer[107-111]. For example, a region upstream of *MYC* (8q24) contains genetic variants that confer increased risk for multiple cancer types, including prostate, breast, colorectal, bladder and chronic lymphocytic leukemia (CLL)[112-115]. This locus contains several functional enhancers, and it was shown that their activity is altered by the cancer-associated SNPs[107-109]. These studies suggest that genetic variation at enhancers and other regulatory elements may be a general feature of susceptibility to cancer and other common diseases.

Cancer is a complex genetic disease that arises from multiple genetic and epigenetic alterations in oncogenes and tumor-suppressor genes[116]. However, different tumor-types are characterized by a set of common hallmarks such as genomic instability and dysregulation of cell cycle[117]. Not surprisingly, cancer cells typically display copy number alterations that affect more than a quarter of their genome[118]. The majority of DNA amplifications involve oncogenes, but they have also been found exclusively in non-coding regions. The increased copy number of an enhancer can amplify its output and cause aberrant gene expression, providing tumor cells with a strong growth advantage. Amplification of non-coding regions seem to be under positive selection since it was observed that they accumulate over time[119], and also that enhancers carrying a risk allele can be preferentially amplified[120,121]. Furthermore, it was shown that non-coding amplifications can specifically affect critical oncogenes (e.g. *MYC*) in different types of tumors[122,123]. The repositioning of enhancers next to oncogenes is a recurrent theme in cancer genomes. This can arise either through large deletions, which frequently occur in carcinomas, or translocations and inversions, which are commonly found in liquid tumors. The latter case is exemplified by chromosomal translocations found in T-cell acute lymphoblastic leukemia (T-ALL) that bring different oncogenes, including *TLX1*, *TLX3*, *TAL1*, *TAL2*, *NOTCH1* and *MYC*, close to the regulatory region of the T-cell receptor[124]. Moreover, large structural rearrangements can affect multiple genes by changing the location of a single regulatory element. Groschel *et al.* showed that the repositioning of an enhancer through inv(3)/t(3;3) underlies the development of AML by deregulating the expression of both *EVI1* and *GATA2*[125]. In addition to large rearrangements, a great number of somatic mutations, involving single-nucleotide alterations, insertions and deletions, are also found in the non-coding cancer genome. However, the identification of non-coding oncogenic mutations is a very challenging task, due to the large size of the non-coding genome, reduced number of whole-genome sequences available, difficulty to assess the function of the mutations and unknown rate background mutation rate. As a consequence, few recurrent mutations in the non-coding genome have been identified so far. Most of these mutations occur in or near promoter regions, such as the ones found upstream of *TERT*[126,127] and *PLEKHS1*[128]. In particular, mutations in the promoter of *TERT* are frequently observed in different types of carcinomas, including bladder, liver, thyroid and melanoma[129-131]. On the contrary, mutations in enhancer elements are expected to be more specific to the tumor-type. This hypothesis is supported by a limited number of cases, such as the mutations that create an enhancer *de novo* in CLL[132] and T-ALL[133]. In addition to the alterations mentioned above, the activity of enhancers can spread locally due to small mutations or deletions that occur in CTCF/cohesin binding sites, which disrupt the boundaries of insulated neighborhoods[57]. Indeed, CTCF binding sites at insulators are among the most altered TF sequences in cancer cells[134] and recent studies have identified recurrent deletions at such boundaries in multiple tumor-types[61]. The finding that proto-oncogenes can be activated through somatic mutations or epigenetic alterations that disrupt CTCF-CTCF loops provides additional evidence for the function of insulated neighborhoods[57,59,61]. Altogether, these studies suggest that the disruption of chromosome architecture, and consequently enhancer activity, contributes to the development of cancer.

CRISPR-Cas systems: from bacterial immunity to genome editing

CRISPR systems were identified in bacteria as an adaptive immune mechanism that protects them from foreign nucleic acids, such as viruses or plasmids[135,136]. Type II-CRISPR systems incorporate invading sequences in the host bacterial genome between an array of repeated sequences. CRISPR repeat arrays are transcribed and processed into CRISPR RNAs (crRNAs), each containing a variable sequence (protospacer sequence) transcribed from the invading genome. A second RNA, known as transactivating CRISPR RNA (tracrRNA), hybridizes with each crRNA and together they form a complex with the Cas9 nuclease[137,138]. The protospacer directs Cas9 to cleave complementary target sequences, provided they are adjacent to a short sequence known as protospacer adjacent motif (PAM). The PAM confers specificity to Cas9 targeting, and enables distinguishing self from non-self DNA sequences[137,138].

The type II CRISPR from *S. pyogenes* was the first system to be engineered for targeted genome editing[137]. The most widely used form of this system is made of two components that must be expressed in cells or organisms to perform DNA editing: the Cas9 nuclease and a single guide RNA (sgRNA), which is a fusion of a crRNA and a fixed tracrRNA. Cas9 can be directed to a specific genomic location by a 20 bp sequence at the 5' end of the sgRNA, which hybridizes with the target sequence by standard RNA-DNA complementary base-pairing rules[137]. The target sites must lie immediately upstream of a canonical PAM sequence (NGG in *S. pyogenes*). Using this system, the Cas9 nuclease can be directed to any DNA sequence of the form N_{20}-NGG simply by changing the first 20 nucleotides of the sgRNA to match the target sequence. Additional CRISPR systems from other bacteria, which recognize alternative PAMs and use different crRNA/tracrRNA sequences, were also adapted for targeted genome editing in human cells[139-141].

The initial demonstration that Cas9 can be programed to cleave DNA *in vitro*[137] propelled a number of studies showing that this platform also functions in a variety of cells and organisms. In 2013, it was shown that Cas9 can target endogenous genes in bacteria[142], immortalized human cell lines[143-146], human pluripotent stem cells[143] and even in a whole organism (*D. rerio*)[146]. The first step for performing targeted genome editing using nucleases is the creation of a DNA double-strand break (DSB) at the target locus[147]. Nuclease-induced DSBs are usually repaired by one of two pathways: non-homologous end joining (NHEJ) or homology-directed repair (HDR). NHEJ is an error-prone repair mechanism that efficiently generates small insertions and deletions (indels) of variable size[148], which can disrupt the coding frame of a gene or the binding site of a TF. HDR-based genome editing can be used to generate specific mutations or insert desired sequences through an exogenous DNA template[149]. The frequency of HDR upon Cas9-mediated DSBs is typically greater than 10% and, in some cases, can reach up to 60%[150]. Given these rates, desired mutations can be simply identified by screening without requiring a drug-resistance selection marker. Cas9 is able to introduce DSBs at multiple sites in parallel, which is a unique advantage of this system compared to other DNA editing tools like ZNFs and TALEs. This strategy has been used to induce large deletions[143], inversions[143,151], and simultaneous mutations in multiple genes[152-154].

CRISPR-Cas9 genome editing has accelerated the generation of cellular and animal transgenic models, expanding biological research beyond genetically tractable model organisms[155]. For example, gene editing can be used to rapidly test the role of specific genetic variants found in the population, instead of relying on animal models that phenocopy a particular disease. This approach was applied in recent years to engineer isogenic ESCs and develop novel transgenic animal models[152,156]. CRISPR-mediated genome editing can also expedite the development of large animal models, including in primates, and thereby accelerate the identification of suitable therapies for humans[156]. CRISPR-Cas9 has also been used for *ex vivo* and *in vivo* gene correction by HDR - either using exogenously supplied oligonucleotides or the endogenous WT allele[157,158]. In the study by Wu and colleagues, it was shown that the resulting mice were fertile and able to transmit the corrected allele to their progeny[158], providing a proof of principle for using CRISPR-Cas9 to correct genetic diseases. There are serious ethical concerns surrounding germline modification of human embryos for correction of disease-causing mutations[159,160]. However, it may be possible to achieve therapeutic benefit for some disorders by correcting faulty genes in somatic cells. This was demonstrated independently by three research groups[161-163] that used CRISPR-Cas9 in a mouse model to delete a mutation that causes Duchenne muscular dystrophy (DMD). This type of approach provides potential means of correcting mutations responsible for DMD and other monogenic disorders[164,165] after birth. As of writing, a number of countries (e.g. Sweden, the UK, Japan and China) have approved research applications based on CRISPR-Cas9 genome editing in human embryos. In the meantime, ongoing clinical trials testing stem cell-based applications[166,167] set the stage for next-generation genome editing therapies. Therefore, it is imperative that health safety investigations keep pace with technological advances of CRISPR systems to ensure an appropriate risk-benefit profile for future therapeutic interventions in human patients.

Genetic screens using CRISPR-Cas9

The simplicity of CRISPR-Cas9 programming has inspired the generation of pooled sgRNA libraries using customized oligonucleotides. Using this approach, a complex pool of oligonucleotides is produced and directly cloned into a plasmid vector to generate a lentiviral library that is used for screening[168]. In 2014, this strategy was successfully employed for both positive and negative forward genetic screens in human and mouse cells[169-171]. These studies revealed both known and novel genes that are involved in fundamental cellular processes and drug/toxin resistance[169-171]. Genome-wide CRISPR screens were also successfully applied *in vivo* to identify protein-coding genes and miRNAs that dictate in cancer progression[172]. Importantly, CRISPR-Cas9 screens display very strong phenotypic effects, likely due to complete knockout of gene expression. Initial comparisons revealed that CRISPR-Cas9 outperformed RNAi in terms of reagent consistency and candidate validation[169]. Recent studies have systematically compared the performance of both technologies in loss-of-function screens[173,174] and it seems that CRISPR has the upper hand in this case (see General discussion).

In pooled CRISPR screens, Cas9 can be either stably expressed in the target cells or encoded in the same lentiviral vector that expresses the sgRNA[175]. The viruses are produced and purified in bulk, and the target cells must be transduced at low multiplicity of infection. This step needs to be optimized in order to avoid cells carrying more than one sgRNA, which can severely compromise the interpretation of the screen. After selection for stable transgene integrations, the mutagenized population of cells undergoes a phenotypic screening in order to identify genes involved in a specific biological process[176]. In positive selection (or enrichment) screens, a strong pressure is applied to select mutations that enhance cellular fitness. This approach is useful to identify genes involved in resistance to toxins[171,177], pathogens[178] and drugs[169,170], but also cellular processes such as metastasis[172]. On the other hand, negative selection (or dropout) screens identify mutations that cause loss of cells during the selection procedure. This type of approach is mainly used to identify essential genes required for cell proliferation and survival[179]. Dropout screens are more sensitive to alterations in the representation of the library because the candidate genes are selected by comparing the abundance of sgRNAs before and after selection. Also, this approach is further complicated by a significant amount of neutral mutations generated by Cas9, which can potentially obscure the desired phenotype[180].

Over the last 30 years, the manipulation of non-coding DNA sequences mainly relied on homologous recombination techniques[181-183]. Recently, this task was greatly facilitated by the development of programmable nucleases, such as ZFNs and TALEs, which can be engineered to cut specific DNA sequences (reviewed in[184]). However, these technologies are low-throughput and, therefore, unsuitable to perform large-scale genetic screens of non-coding DNA sequences. The advent of CRISPR systems filled a technological gap and, not surprisingly, they were applied in forward genetic screens of non-coding DNA elements[185,186]. These studies identified novel enhancers and other regulatory elements involved in oncogene-induced senescence[185], cancer cell growth[185] and drug resistance[186]. Remarkably, it was observed that mutations in enhancers cause phenotypic effects comparable to that of mutations in their target genes[185,186]. These results emphasize the importance of identifying causal non-coding variants that contribute to the development of human diseases (see General discussion). To date, genetic screens of non-coding DNA sequences have been confined to mutagenesis over regions of 2 kb to 1 Mb[187]. Given the fast pace of technological advances, it is safe to say that CRISPR-Cas9 screens are destined to generate an immense amount of data and contribute decisively to elucidate *all* the functions of the human genome.

References

1 Ptashne, M. Regulation of transcription: from lambda to eukaryotes. *Trends Biochem Sci* **30**, 275-279, doi:10.1016/j.tibs.2005.04.003 (2005).

2 Pertea, M. & Salzberg, S. L. Between a chicken and a grape: estimating the number of human genes. *Genome Biol* **11**, 206, doi:10.1186/gb-2010-11-5-206 (2010).

3 Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299-308 (1981).

4 Amano, T. *et al.* Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev Cell* **16**, 47-57, doi:10.1016/j.devcel.2008.11.011 (2009).

5 Shi, J. *et al.* Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. *Genes Dev* **27**, 2648-2662, doi:10.1101/gad.232710.113 (2013).

6 Levine, M. Transcriptional enhancers in animal development and evolution. *Curr Biol* **20**, R754-763, doi:10.1016/j.cub.2010.06.070 (2010).

7 Moreau, P. *et al.* The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res* **9**, 6047-6068 (1981).

8 Odom, D. T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39**, 730-732, doi:10.1038/ng2047 (2007).

9 Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554-566, doi:10.1016/j.cell.2015.01.006 (2015).

10 Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858, doi:10.1038/nature07730 (2009).

11 Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199-205, doi:10.1038/nature08451 (2009).

12 Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187, doi:10.1038/nature09033 (2010).

13 De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol* **8**, e1000384, doi:10.1371/journal.pbio.1000384 (2010).

14 Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* **57**, 159-197, doi:10.1146/annurev.bi.57.070188.001111 (1988).

15 Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318, doi:10.1038/ng1966 (2007).

16 Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936, doi:10.1073/pnas.1016071107 (2010).

17 Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283, doi:10.1038/nature09692 (2011).

18 Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-112, doi:10.1038/nature07829 (2009).

19 Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49, doi:10.1038/nature09906 (2011).

20 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

21 Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82, doi:10.1038/nature11232 (2012).

22 Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327-339, doi:10.1016/j.cell.2011.01.024 (2011).

23 Bulger, M. & Groudine, M. Looping versus linking: toward a model for long-distance gene activation. *Genes Dev* **13**, 2465-2477 (1999).

24 Blackwood, E. M. & Kadonaga, J. T. Going the distance: a current view of enhancer action. *Science* **281**, 60-63 (1998).

25 Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-1311, doi:10.1126/science.1067799 (2002).

26 Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* **38**, 1341-1347, doi:10.1038/ng1891 (2006).

27 Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* **16**, 1299-1309, doi:10.1101/gr.5571506 (2006).

28 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

29 de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev* **26**, 11-24, doi:10.1101/gad.179804.111 (2012).

30 Drissen, R. *et al.* The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes Dev* **18**, 2485-2490, doi:10.1101/gad.317004 (2004).

31 Vakoc, C. R. *et al.* Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Mol Cell* **17**, 453-462, doi:10.1016/j.molcel.2004.12.028 (2005).

32 Ohtsuki, S. & Levine, M. GAGA mediates the enhancer blocking activity of the eve promoter in the Drosophila embryo. *Genes Dev* **12**, 3325-3330 (1998).

33 Butler, J. E. & Kadonaga, J. T. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev* **15**, 2515-2519, doi:10.1101/gad.924301 (2001).

34 Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233-1244, doi:10.1016/j.cell.2012.03.051 (2012).

35 Deng, W. *et al.* Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell* **158**, 849-860, doi:10.1016/j.cell.2014.05.050 (2014).

36 Cremer, T. & Cremer, M. Chromosome territories. *Cold Spring Harb Perspect Biol* **2**, a003889, doi:10.1101/cshperspect.a003889 (2010).

37 Gibcus, J. H. & Dekker, J. The hierarchy of the 3D genome. *Mol Cell* **49**, 773-782, doi:10.1016/j.molcel.2013.02.011 (2013).

38 Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294, doi:10.1038/nature12644 (2013).

39 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

40 Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-385, doi:10.1038/nature11049 (2012).

41 Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336, doi:10.1038/nature14222 (2015).

42 Iborra, F. J., Pombo, A., Jackson, D. A. & Cook, P. R. Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. *J Cell Sci* **109 (Pt 6)**, 1427-1436 (1996).

43 Sutherland, H. & Bickmore, W. A. Transcription factories: gene expression in unions? *Nat Rev Genet* **10**, 457-466, doi:10.1038/nrg2592 (2009).

44 Bell, A. C., West, A. G. & Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**, 387-396 (1999).

45 Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**, 1194-1211, doi:10.1016/j.cell.2009.06.001 (2009).

46 Yusufzai, T. M., Tagami, H., Nakatani, Y. & Felsenfeld, G. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol Cell* **13**, 291-298 (2004).

47 Splinter, E. *et al.* CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev* **20**, 2349-2354, doi:10.1101/gad.399506 (2006).

48 Hou, C., Dale, R. & Dean, A. Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc Natl Acad Sci U S A* **107**, 3651-3656, doi:10.1073/pnas.0912087107 (2010).

49 Rubio, E. D. *et al.* CTCF physically links cohesin to chromatin. *Proc Natl Acad Sci U S A* **105**, 8309-8314, doi:10.1073/pnas.0801273105 (2008).

50 Wendt, K. S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**, 796-801, doi:10.1038/nature06634 (2008).

51 Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430-435, doi:10.1038/nature09380 (2010).

52 Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58-64, doi:10.1038/nature08497 (2009).

53 DeMare, L. E. *et al.* The genomic landscape of cohesin-associated chromatin interactions. *Genome Res* **23**, 1224-1234, doi:10.1101/gr.156570.113 (2013).

54 Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281-1295, doi:10.1016/j.cell.2013.04.053 (2013).

55 Dowen, J. M. *et al.* Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374-387, doi:10.1016/j.cell.2014.09.030 (2014).

56 Hnisz, D., Day, D. S. & Young, R. A. Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell* **167**, 1188-1200, doi:10.1016/j.cell.2016.10.024 (2016).

57 Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454-1458, doi:10.1126/science.aad9024 (2016).

58 Ji, X. *et al.* 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell Stem Cell* **18**, 262-275, doi:10.1016/j.stem.2015.11.007 (2016).

59 Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110-114, doi:10.1038/nature16490 (2016).

60 Narendra, V. *et al.* CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* **347**, 1017-1021, doi:10.1126/science.1262088 (2015).

61 Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet* **47**, 818-821, doi:10.1038/ng.3335 (2015).

62 Fukaya, T., Lim, B. & Levine, M. Enhancer Control of Transcriptional Bursting. *Cell* **166**, 358-368, doi:10.1016/j.cell.2016.05.025 (2016).

63 Harris, H. Turnover of nuclear and cytoplasmic ribonucleic acid in two types of animal cell, with some further observations on the nucleolus. *Biochem J* **73**, 362-369 (1959).

64 Collis, P., Antoniou, M. & Grosveld, F. Definition of the minimal requirements within the human beta-globin gene and the dominant control region for high level expression. *EMBO J* **9**, 233-240 (1990).

65 Tuan, D., Kong, S. & Hu, K. Transcription of the hypersensitive site HS2 enhancer in erythroid cells. *Proc Natl Acad Sci U S A* **89**, 11219-11223 (1992).

66 Ashe, H. L., Monks, J., Wijgerde, M., Fraser, P. & Proudfoot, N. J. Intergenic transcription and transinduction of the human beta-globin locus. *Genes Dev* **11**, 2494-2509 (1997).

67 Masternak, K., Peyraud, N., Krawczyk, M., Barras, E. & Reith, W. Chromatin remodeling and extragenic transcription at the MHC class II locus control region. *Nat Immunol* **4**, 132-137, doi:10.1038/ni883 (2003).

68 Ho, Y., Elefant, F., Liebhaber, S. A. & Cooke, N. E. Locus control region transcription plays an active role in long-range gene activation. *Mol Cell* **23**, 365-375, doi:10.1016/j.molcel.2006.05.041 (2006).

69 Koch, F. *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat Struct Mol Biol* **18**, 956-963, doi:10.1038/nsmb.2085 (2011).

70 Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108, doi:10.1038/nature11233 (2012).

71 Hah, N. *et al.* A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145**, 622-634, doi:10.1016/j.cell.2011.03.042 (2011).

72 Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474**, 390-394, doi:10.1038/nature10006 (2011).

73 Kaikkonen, M. U. *et al.* Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Mol Cell* **51**, 310-325, doi:10.1016/j.molcel.2013.07.010 (2013).

74 Li, W. *et al.* Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516-520, doi:10.1038/nature12210 (2013).

75 Allen, M. A. *et al.* Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. *Elife* **3**, e02200, doi:10.7554/eLife.02200 (2014).

76 Leveille, N. *et al.* Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA. *Nat Commun* **6**, 6520, doi:10.1038/ncomms7520 (2015).

77 Hah, N., Murakami, S., Nagari, A., Danko, C. G. & Kraus, W. L. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res* **23**, 1210-1223, doi:10.1101/gr.152306.112 (2013).

78 Melo, C. A. *et al.* eRNAs are required for p53-dependent enhancer activity and gene transcription. *Mol Cell* **49**, 524-535, doi:10.1016/j.molcel.2012.11.021 (2013).

79 Orom, U. A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58, doi:10.1016/j.cell.2010.09.001 (2010).

80 Lai, F. *et al.* Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* **494**, 497-501, doi:10.1038/nature11884 (2013).

81 Lam, M. T. *et al.* Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature* **498**, 511-515, doi:10.1038/nature12209 (2013).

82 Gribnau, J., Diderich, K., Pruzina, S., Calzolari, R. & Fraser, P. Intergenic transcription and developmental remodeling of chromatin subdomains in the human beta-globin locus. *Mol Cell* **5**, 377-386 (2000).

83 Mousavi, K. *et al.* eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol Cell* **51**, 606-617, doi:10.1016/j.molcel.2013.07.022 (2013).

84 Johnson, K. D. *et al.* Highly restricted localization of RNA polymerase II within a locus control region of a tissue-specific chromatin domain. *Mol Cell Biol* **23**, 6484-6493 (2003).

85 Lin, Y. C. *et al.* Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nat Immunol* **13**, 1196-1204, doi:10.1038/ni.2432 (2012).

86 Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113, doi:10.1038/nature11279 (2012).

87 Schaukowitch, K. *et al.* Enhancer RNA facilitates NELF release from immediate early genes. *Mol Cell* **56**, 29-42, doi:10.1016/j.molcel.2014.08.023 (2014).

88 Sur, I. & Taipale, J. The role of enhancers in cancer. *Nat Rev Cancer* **16**, 483-493, doi:10.1038/nrc.2016.62 (2016).

89 Vousden, K. H. & Prives, C. Blinded by the Light: The Growing Complexity of p53. *Cell* **137**, 413-431, doi:10.1016/j.cell.2009.04.037 (2009).

90 Petitjean, A., Achatz, M. I., Borresen-Dale, A. L., Hainaut, P. & Olivier, M. TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene* **26**, 2157-2165, doi:10.1038/sj.onc.1210302 (2007).

91 Hayashi, S. I. *et al.* The expression and function of estrogen receptor alpha and beta in human breast cancer and its clinical application. *Endocr Relat Cancer* **10**, 193-202 (2003).

92 Carroll, J. S. *et al.* Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* **38**, 1289-1297, doi:10.1038/ng1901 (2006).

93 Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**, 958-970, doi:10.1016/j.cell.2008.01.018 (2008).

94 Beelen, K., Zwart, W. & Linn, S. C. Can predictive biomarkers in breast cancer guide adjuvant endocrine therapy? *Nat Rev Clin Oncol* **9**, 529-541, doi:10.1038/nrclinonc.2012.121 (2012).

95 Shi, J. & Vakoc, C. R. The mechanisms behind the therapeutic activity of BET bromodomain inhibition. *Mol Cell* **54**, 728-736, doi:10.1016/j.molcel.2014.05.016 (2014).

96 Miguel-Escalada, I., Pasquali, L. & Ferrer, J. Transcriptional enhancers: functional insights and role in human disease. *Curr Opin Genet Dev* **33**, 71-76, doi:10.1016/j.gde.2015.08.009 (2015).

97 Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**, 1725-1735 (2003).

98 Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M. & Shiroishi, T. Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* **132**, 797-803, doi:10.1242/dev.01613 (2005).

99 Benko, S. *et al.* Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nat Genet* **41**, 359-364, doi:10.1038/ng.329 (2009).

100 Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum Mol Genet* **21**, 3255-3263, doi:10.1093/hmg/dds165 (2012).

101 Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-1006, doi:10.1093/nar/gkt1229 (2014).

102 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-9367, doi:10.1073/pnas.0903103106 (2009).

103 Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* **363**, 166-176, doi:10.1056/NEJMra0905980 (2010).

104 Li, Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633-641, doi:10.1016/j.cell.2012.12.034 (2013).

105 Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* **16**, 197-212, doi:10.1038/nrg3891 (2015).

106 Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461, doi:10.1038/nature12787 (2014).

107 Jia, L. *et al.* Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet* **5**, e1000597, doi:10.1371/journal.pgen.1000597 (2009).

108 Tuupanen, S. *et al.* The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* **41**, 885-890, doi:10.1038/ng.406 (2009).

109 Pomerantz, M. M. *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* **41**, 882-884, doi:10.1038/ng.403 (2009).

110 Oldridge, D. A. *et al.* Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism. *Nature* **528**, 418-421, doi:10.1038/nature15540 (2015).

111 Dunning, A. M. *et al.* Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. *Nat Genet* **48**, 374-386, doi:10.1038/ng.3521 (2016).

112 Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* **39**, 631-637, doi:10.1038/ng1999 (2007).

113 Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* **39**, 984-988, doi:10.1038/ng2085 (2007).

114 Ghoussaini, M. *et al.* Multiple loci with different cancer specificities within the 8q24 gene desert. *J Natl Cancer Inst* **100**, 962-966, doi:10.1093/jnci/djn190 (2008).

115 Crowther-Swanepoel, D. *et al.* Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nat Genet* **42**, 132-136, doi:10.1038/ng.510 (2010).

116 Kinzler, K. W. & Vogelstein, B. Lessons from hereditary colorectal cancer. *Cell* **87**, 159-170 (1996).

117 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).

118 Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899-905, doi:10.1038/nature08822 (2010).

119 Hsu, P. Y. *et al.* Amplification of distant estrogen response elements deregulates target genes associated with tamoxifen resistance in breast cancer. *Cancer Cell* **24**, 197-212, doi:10.1016/j.ccr.2013.07.007 (2013).

120 Tuupanen, S. *et al.* Allelic imbalance at rs6983267 suggests selection of the risk allele in somatic colorectal tumor evolution. *Cancer Res* **68**, 14-17, doi:10.1158/0008-5472.CAN-07-5766 (2008).

121 Sur, I. K. *et al.* Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* **338**, 1360-1363, doi:10.1126/science.1228606 (2012).

122 Herranz, D. *et al.* A NOTCH1-driven MYC enhancer promotes T cell development, transformation and acute lymphoblastic leukemia. *Nat Med* **20**, 1130-1137, doi:10.1038/nm.3665 (2014).

123 Zhang, X. *et al.* Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat Genet* **48**, 176-182, doi:10.1038/ng.3470 (2016).

124 Cauwelier, B. *et al.* Molecular cytogenetic study of 126 unselected T-ALL cases reveals high incidence of TCRbeta locus rearrangements and putative new T-cell oncogenes. *Leukemia* **20**, 1238-1244, doi:10.1038/sj.leu.2404243 (2006).

125 Groschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**, 369-381, doi:10.1016/j.cell.2014.02.019 (2014).

126 Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959-961, doi:10.1126/science.1230062 (2013).

127 Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957-959, doi:10.1126/science.1229259 (2013).

128 Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* **46**, 1160-1165, doi:10.1038/ng.3101 (2014).

129 Kinde, I. *et al.* TERT promoter mutations occur early in urothelial neoplasia and are biomarkers of early disease and disease recurrence in urine. *Cancer Res* **73**, 7162-7167, doi:10.1158/0008-5472.CAN-13-2498 (2013).

130 Nault, J. C. *et al.* High frequency of telomerase reverse-transcriptase promoter somatic mutations in hepatocellular carcinoma and preneoplastic lesions. *Nat Commun* **4**, 2218, doi:10.1038/ncomms3218 (2013).

131 Vinagre, J. *et al.* Frequency of TERT promoter mutations in human cancers. *Nat Commun* **4**, 2185, doi:10.1038/ncomms3185 (2013).

132 Puente, X. S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519-524, doi:10.1038/nature14666 (2015).

133 Mansour, M. R. *et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373-1377, doi:10.1126/science.1259037 (2014).

134 Kaiser, V. B., Taylor, M. S. & Semple, C. A. Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLoS Genet* **12**, e1006207, doi:10.1371/journal.pgen.1006207 (2016).

135 Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167-170, doi:10.1126/science.1179555 (2010).

136 Wiedenheft, B., Sternberg, S. H. & Doudna, J. A. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482**, 331-338, doi:10.1038/nature10886 (2012).

137 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821, doi:10.1126/science.1225829 (2012).

138 Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602-607, doi:10.1038/nature09886 (2011).

139 Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci U S A* **109**, E2579-2586, doi:10.1073/pnas.1208507109 (2012).

140 Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat Methods* **10**, 1116-1121, doi:10.1038/nmeth.2681 (2013).

141 Hou, Z. *et al.* Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis. *Proc Natl Acad Sci U S A* **110**, 15644-15649, doi:10.1073/pnas.1313587110 (2013).

142 Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat Biotechnol* **31**, 233-239, doi:10.1038/nbt.2508 (2013).

143 Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823-826, doi:10.1126/science.1232033 (2013).

144 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823, doi:10.1126/science.1231143 (2013).

145 Jinek, M. *et al.* RNA-programmed genome editing in human cells. *Elife* **2**, e00471, doi:10.7554/eLife.00471 (2013).

146 Hwang, W. Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat Biotechnol* **31**, 227-229, doi:10.1038/nbt.2501 (2013).

147 Carroll, D. Genome engineering with zinc-finger nucleases. *Genetics* **188**, 773-782, doi:10.1534/genetics.111.131433 (2011).

148 Lieber, M. R. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu Rev Biochem* **79**, 181-211, doi:10.1146/annurev.biochem.052308.093131 (2010).

149 Jasin, M. & Rothstein, R. Repair of strand breaks by homologous recombination. *Cold Spring Harb Perspect Biol* **5**, a012740, doi:10.1101/cshperspect.a012740 (2013).

150 Richardson, C. D., Ray, G. J., DeWitt, M. A., Curie, G. L. & Corn, J. E. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat Biotechnol* **34**, 339-344, doi:10.1038/nbt.3481 (2016).

151 Maddalo, D. *et al.* In vivo engineering of oncogenic chromosomal rearrangements with the CRISPR/Cas9 system. *Nature* **516**, 423-427, doi:10.1038/nature13902 (2014).

152 Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910-918, doi:10.1016/j.cell.2013.04.025 (2013).

153 Li, W., Teng, F., Li, T. & Zhou, Q. Simultaneous generation and germline transmission of multiple gene mutations in rat using CRISPR-Cas systems. *Nat Biotechnol* **31**, 684-686, doi:10.1038/nbt.2652 (2013).

154 Jao, L. E., Wente, S. R. & Chen, W. Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proc Natl Acad Sci U S A* **110**, 13904-13909, doi:10.1073/pnas.1308335110 (2013).

155 Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol* **32**, 347-355, doi:10.1038/nbt.2842 (2014).

156 Niu, Y. *et al.* Generation of gene-modified cynomolgus monkey via Cas9/RNA-mediated gene targeting in one-cell embryos. *Cell* **156**, 836-843, doi:10.1016/j.cell.2014.01.027 (2014).

157 Schwank, G. *et al.* Functional repair of CFTR by CRISPR/Cas9 in intestinal stem cell organoids of cystic fibrosis patients. *Cell Stem Cell* **13**, 653-658, doi:10.1016/j.stem.2013.11.002 (2013).

158 Wu, Y. *et al.* Correction of a genetic disease in mouse via use of CRISPR-Cas9. *Cell Stem Cell* **13**, 659-662, doi:10.1016/j.stem.2013.10.016 (2013).

159 Bosley, K. S. *et al.* CRISPR germline engineering--the community speaks. *Nat Biotechnol* **33**, 478-486, doi:10.1038/nbt.3227 (2015).

160 Baltimore, D. *et al.* Biotechnology. A prudent path forward for genomic engineering and germline gene modification. *Science* **348**, 36-38, doi:10.1126/science.aab1028 (2015).

161 Long, C. *et al.* Postnatal genome editing partially restores dystrophin expression in a mouse model of muscular dystrophy. *Science* **351**, 400-403, doi:10.1126/science.aad5725 (2016).

162 Nelson, C. E. *et al.* In vivo genome editing improves muscle function in a mouse model of Duchenne muscular dystrophy. *Science* **351**, 403-407, doi:10.1126/science.aad5143 (2016).

163 Tabebordbar, M. *et al.* In vivo gene editing in dystrophic mouse muscle and muscle stem cells. *Science* **351**, 407-411, doi:10.1126/science.aad5177 (2016).

164 Xie, F. *et al.* Seamless gene correction of beta-thalassemia mutations in patient-specific iPSCs using CRISPR/Cas9 and piggyBac. *Genome Res* **24**, 1526-1533, doi:10.1101/gr.173427.114 (2014).

165 Song, B. *et al.* Improved hematopoietic differentiation efficiency of gene-corrected beta-thalassemia induced pluripotent stem cells by CRISPR/Cas9 system. *Stem Cells Dev* **24**, 1053-1065, doi:10.1089/scd.2014.0347 (2015).

166 Kamao, H. *et al.* Characterization of human induced pluripotent stem cell-derived retinal pigment epithelium cell sheets aiming for clinical application. *Stem Cell Reports* **2**, 205-218, doi:10.1016/j.stemcr.2013.12.007 (2014).

167 Hotta, A. & Yamanaka, S. From Genomics to Gene Therapy: Induced Pluripotent Stem Cells Meet Genome Editing. *Annu Rev Genet* **49**, 47-70, doi:10.1146/annurev-genet-112414-054926 (2015).

168 Root, D. E., Hacohen, N., Hahn, W. C., Lander, E. S. & Sabatini, D. M. Genome-scale loss-of-function screening with a lentiviral RNAi library. *Nat Methods* **3**, 715-719, doi:10.1038/nmeth924 (2006).

169 Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87, doi:10.1126/science.1247005 (2014).

170 Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84, doi:10.1126/science.1246981 (2014).

171 Koike-Yusa, H., Li, Y., Tan, E. P., Velasco-Herrera Mdel, C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat Biotechnol* **32**, 267-273, doi:10.1038/nbt.2800 (2014).

172 Chen, S. *et al.* Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell* **160**, 1246-1260, doi:10.1016/j.cell.2015.02.038 (2015).

173 Evers, B. *et al.* CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. *Nat Biotechnol* **34**, 631-633, doi:10.1038/nbt.3536 (2016).

174 Morgens, D. W., Deans, R. M., Li, A. & Bassik, M. C. Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes. *Nat Biotechnol* **34**, 634-636, doi:10.1038/nbt.3567 (2016).

175 Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods* **11**, 783-784, doi:10.1038/nmeth.3047 (2014).

176 Boutros, M. & Ahringer, J. The art and design of genetic screens: RNA interference. *Nat Rev Genet* **9**, 554-566, doi:10.1038/nrg2364 (2008).

177 Zhou, Y. *et al.* High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature* **509**, 487-491, doi:10.1038/nature13166 (2014).

178 Parnas, O. *et al.* A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. *Cell* **162**, 675-686, doi:10.1016/j.cell.2015.06.059 (2015).

179 Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-1101, doi:10.1126/science.aac7041 (2015).

180 Shalem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR-Cas9. *Nat Rev Genet* **16**, 299-311, doi:10.1038/nrg3899 (2015).

181 Smithies, O., Gregg, R. G., Boggs, S. S., Koralewski, M. A. & Kucherlapati, R. S. Insertion of DNA sequences into the human chromosomal beta-globin locus by homologous recombination. *Nature* **317**, 230-234 (1985).

182 Bender, M. A. *et al.* Description and targeted deletion of 5' hypersensitive site 5 and 6 of the mouse beta-globin locus control region. *Blood* **92**, 4394-4403 (1998).

183 Reik, A. *et al.* The locus control region is necessary for gene expression in the human beta-globin locus but not the maintenance of an open chromatin structure in erythroid cells. *Mol Cell Biol* **18**, 5992-6000 (1998).

184 Hilton, I. B. & Gersbach, C. A. Enabling functional genomics with genome engineering. *Genome Res* **25**, 1442-1455, doi:10.1101/gr.190124.115 (2015).

185 Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol* **34**, 192-198, doi:10.1038/nbt.3450 (2016).

186 Sanjana, N. E. *et al.* High-resolution interrogation of functional elements in the noncoding genome. *Science* **353**, 1545-1549, doi:10.1126/science.aaf7613 (2016).

187 Lopes, R., Korkmaz, G. & Agami, R. Applying CRISPR-Cas9 tools to identify and characterize transcriptional enhancers. *Nat Rev Mol Cell Biol* **17**, 597-604, doi:10.1038/nrm.2016.79 (2016).

Chapter 2

Applying CRISPR-Cas9 tools to identify and characterize transcriptional enhancers

Adapted from Nature Reviews Molecular Cell Biology. 2016 Sep;17(9):597-604.

Applying CRISPR–Cas9 tools to identify and characterize transcriptional enhancers

Rui Lopes, Gozde Korkmaz and Reuven Agami

Abstract | The development of the CRISPR–Cas9 system triggered a revolution in the field of genome engineering. Initially, the use of this system was focused on the study of protein-coding genes but, recently, a number of CRISPR–Cas9-based tools have been developed to study non-coding transcriptional regulatory elements. These technological advances offer unprecedented opportunities for elucidating the functions of enhancers in their endogenous context. Here, we discuss the application, current limitations and future development of CRISPR–Cas9 systems to identify and characterize enhancer elements in a high-throughput manner.

Enhancers were initially identified as distal regulatory elements that increase transcription independently of their orientation, position and distance relative to a target promoter[1]. These elements are essential for precise spatiotemporal regulation of gene expression, which is required for proper cell development, differentiation and homeostasis[2]. A key feature of enhancers is their ability to function as transcription factor binding platforms, and genetic alterations at these regions are associated with pathological states[3–8].

Enhancers are usually defined by criteria unrelated to their endogenous biological function, such as the makeup of specific histone modifications, increased chromatin accessibility and the occurrence of bidirectional transcription in the loci in question[9–11]. Based on these features, several techniques have been used to identify putative enhancers and validate their activity in a high-throughput fashion (BOX 1). Nonetheless, these methods cannot fully determine whether a putative enhancer is required for transcription in its native context or which gene (or genes) it regulates.

The emergence of the CRISPR–Cas9 technology has opened unprecedented opportunities for targeted genome editing in human cells[12–15]. The RNA-guided endonuclease, Cas9 can be targeted to any genomic locus by sequence-specific single guide RNA (sgRNA) to initiate double-stranded DNA cleavage (BOX 2). Importantly, Cas9 can be directed to non-coding transcriptional regulatory elements, such as promoters[16,17] and enhancers[18,19]. This feature led to the development of several Cas9-based methods and tools to study non-coding regulatory elements in their native chromatin environment. Here, we summarize recent advances in CRISPR–Cas9 systems that allow the modulation of the activity of enhancers by altering their transcriptional, epigenetic or genetic features, and discuss the current limitations of these technologies in the study of enhancers. We also provide a perspective on future developments of CRISPR–Cas9 tools and their application in identifying and characterizing enhancers in an unbiased fashion.

Transcriptional modulation of enhancers

An attractive approach to studying enhancers in their endogenous context is the use of CRISPR–Cas9 tools to enforce enhancer activation or repression directly.

Activating enhancer elements. CRISPR-mediated gene activation (CRISPRa) uses nuclease-deactivated Cas9 (known as dCas9) fused to transcription activating domains (BOX 2). This approach was first tested on promoters using the transcription activation domains of the VP64 (a tetrameric repeat of herpes simplex virus protein VP16) and p65 (nuclear factor-κB p65 subunit) proteins. Indeed, dCas9–VP64 and dCas9–p65 could activate endogenous genes when targeted by one sgRNA to their promoter region[16,20,21]. Yet, each protein fusion required the combination of several sgRNAs per gene to achieve high expression levels of the target gene[21,22]. To overcome this limitation, several groups developed CRISPRa tools containing multiple activation domains[23,24]. These improvements made the simultaneous activation of dozens of genes as well as genome-wide activation screens possible.

The activating capacity of VP64 relies on the sequential recruitment of cofactors, such as the histone acetyltransferase p300 and TFIID[25,26], suggesting that it can potentially be used to modulate enhancer activity. Indeed, a dCas9–VP64 fusion was shown to activate endogenous genes by recruiting p300 to acetylate H3K27 (histone H3 Lys27) at their cognate enhancer regions[27]. However, targeting dCas9–VP64 to the locus control region (LCR) of the haemoglobin genes failed to activate some of them; dCas9–VP64 was able to induce H3K27 acetylation at the targeted enhancers but not at the cognate haemoglobin promoters, and this correlated with lack of activation of gene expression[28]. This suggests that the recruitment of transcription pre-initiation complex (PIC) components by VP64 (REFS 25,26) might not be sufficient to elicit enhancer activity. It is therefore not clear to what extent dCas9–VP64 can activate enhancer elements.

Repressing enhancer elements. The utility of CRISPR–Cas9-based tools in repressing gene expression (termed CRISPR interference (CRISPRi)) was first demonstrated by showing that a dCas9–sgRNA complex could block RNA polymerase elongation[29]. This strategy was also successfully applied to interfere with transcription factor binding, thereby disrupting enhancer activity[27].

The use of dCas9 alone achieved only modest repression of gene expression in mammalian cells. To improve CRISPRi, dCas9 was fused with repressive effectors, such as the Krüppel-associated box (KRAB) domain of KOX1 (REFS 16,30). A dCas9–KRAB fusion was shown to efficiently repress the expression of protein-coding and non-coding genes on a genome-wide scale[17]. KRAB fusions were later shown to inactivate the expression of endogenous genes by targeting their distal enhancer elements[27,31]. Indeed, dCas9–KRAB can be directed to single elements of composite enhancers and still achieve highly specific repression of genes[32,33]. This effect seems to be mediated by decreased chromatin accessibility at both enhancer and target promoters[32]. It was also noted that directing dCas9–KRAB to enhancer regions resulted in higher H3K9me3 (histone H3 Lys9 trimethylation) levels at the target promoter[32,34]. Therefore, it is possible that KRAB fusions generate off-target effects through heterochromatin spreading[35], and that they silence promoter activity rather than inactivating the target enhancer.

Box 1 | Methods for genome-wide identification of putative gene regulatory elements

The comprehensive identification of transcriptional regulatory elements is a major challenge in genomic research. The emergence of next-generation sequencing propelled the development of a number of high-throughput methods that were used to identify putative enhancers based on their features and activity.

Transcription factor binding sites are the core building blocks of regulatory elements. ChIP–seq (chromatin immunoprecipitation followed by sequencing) is the most common technique to determine transcription-factor occupancy on a genome-wide scale[72]. This method is particularly suitable for identifying stimulus-induced changes or occupancy across different cell types, as well as for identifying binding sites of transcription cofactors (such as the histone acetyltransferase p300), which are frequently associated with enhancers[73]. Genome-wide mapping of histone post-translational modifications by ChIP–seq revealed that H3K4me1 (histone H3 Lys4 monomethylation) and H3K27ac (H3K27 acetylation) are enriched at active-enhancer regions[74], whereas active-promoter regions are marked by H3K4me3 and H3K27ac. This 'histone code' is widely used to annotate regulatory elements and corresponds well with heterologous reporter assays[60], which have been commonly used to interrogate the activity of any genomic region. Cloning of a putative enhancer downstream or upstream of a reporter gene can reflect its functionality by inducing the expression of the reporter gene.

Active-enhancer regions are depleted of nucleosomes, leaving the DNA accessible to enzymatic cleavage. This has been exploited to identify regulatory regions across the genome, for example by coupling digestion by DNase I with DNase-seq (DNase I hypersensitive sites sequencing)[75].

Distal enhancers are brought into spatial proximity with their target promoters through DNA looping. Chromosome conformation capture (3C)[76] and its variants[77] are useful for predicting putative enhancers and their target genes. Chromatin interaction analysis with paired-end tag (ChIA–PET) sequencing[78] is a combination of 3C-based methods with ChIP, which enables the identification of proteins involved in the formation of specific chromosomal contacts.

Active enhancers support divergent transcription of their own loci[79]. The expression of these enhancer-associated RNAs (eRNAs) has been used to identify active enhancers in a cell type-specific manner, because eRNA expression correlates well with enhancer activity. eRNA expression can be detected by methods that measure nascent RNAs, such as global run-on sequencing[80] (GRO-seq).

The ability to increase transcription from minimal promoters in heterologous reporter vectors is a hallmark of enhancer activity[1], and has been used to identify enhancer elements both in cells and in animals. Initially, these methods had low throughput. The recent development of massive parallel reporter assays[81,82] (MPRA) and self-transcribing active regulatory region sequencing[83] (STARR-seq) enables the evaluation of enhancer activity of thousands or even millions of DNA sequences simultaneously.

The combined application of the techniques mentioned above led to a tremendous increase in our understanding of transcriptional regulatory elements. DNase-seq in combination with ChIP–seq of epigenetic modifications was applied to predict putative enhancer regions. Yet, the epigenetic status of a region is not identical to enhancer activity, mostly owing to the use of arbitrary cut-offs of histone modifications ratios (for example, H3K4me1/H3K4me3) as a measure of activity[84–86], which resulted in the estimation that the human genome harbours approximately one million enhancers[84–86]. Instead, when methods that directly measure transcriptional activity were used[87], only 40,000 to 65,000 transcriptionally active putative enhancers were predicted (though any active but not transcribed enhancer would have been missed)[9,88].

Altogether, these techniques cannot assess specific enhancer functionality because they provide only circumstantial evidence that the identified elements are actively engaged in transcription regulation. Thus, other methods are required for the identification of enhancers that drive gene expression in certain biological settings.

Epigenetic modulation of enhancers

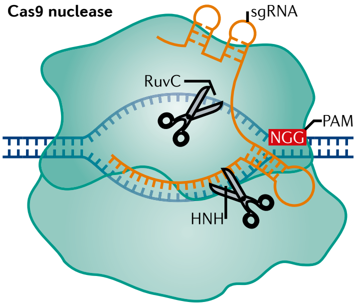
As noted above, enhancer activity is associated with dynamic epigenetic states, including acetylation and methylation of histone tails[36,37]. The ability to modulate these epigenetic modifications is essential to gain better understanding of their importance for enhancer function. This prompted the development of several epigenome-editing tools that are based on fusions of epigenetic regulators (writers and erasers) to dCas9 and allow direct manipulation of epigenetic states of gene regulatory elements.

Lys-specific histone demethylase 1 (LSD1) catalyses the removal of H3K4 methylation[38]. Recently, a dCas9–LSD1 fusion was targeted to a distal enhancer of the pluripotency gene *Oct4* (also known as *Pou5f1*) and achieved specific gene repression and loss of pluripotency in embryonic stem (ES) cells[34]. It was also shown that enhancer targeting by dCas9–LSD1 caused a marked decrease in H3K4me2 and H3K27Ac marks[34,38], consistent with reduced enhancer activity. However, dCas9–LSD1 was unable to repress genes when targeted to promoter regions[34], a surprising result given the well-documented role of LSD1 in the repression of endogenous promoters[38]. A possible explanation for these results is that LSD1 alone is ineffective and may require additional cofactors to inactivate some regulatory elements, indicating that current LSD1 fusions may have low efficacy, compromising their application in high-throughput screens of enhancer elements.

A fusion protein of dCas9 with the catalytic core domain of p300 (dCas9–p300) was created recently[28], and this allowed the manipulation of H3K27ac levels at both proximal and distal regulatory elements. Interestingly, dCas9–p300 was capable of activating genes with high specificity using one sgRNA per target gene[28] and with higher transactivation capacity than dCas9-VP64 (REF. 28). This was particularly evident at distal enhancer elements, where dCas9–VP64 displayed little capacity to activate target genes[28]. These results establish dCas9–p300 as a robust tool to modulate histone acetylation and activate gene expression. However, it is unclear whether dCas9–p300 is suitable for use in genome-wide functional screens. The higher activity of dCas9–p300 seems to be related to the direct acetylation of downstream target promoters, as opposed to the indirect recruitment of PIC by VP64 (REFS 25,26). Still, it remains to be determined whether acetylation of H3K27 alone is sufficient for gene activation.

The consequences of manipulating epigenetic states on transcription regulation are not completely understood, and the utilization of epigenome-editing tools might be associated with potential off-target effects. Nevertheless, the expansion of the current dCas9 epigenetic toolbox will be much needed to shed light on the connections between the epigenome and gene expression. Moreover, the precise manipulation of epigenetic modifications holds tremendous therapeutic potential.

Box 2 | Genome editing and regulation using CRISPR–Cas9 systems



The CRISPR–Cas9 system is based on the molecular machinery of type II CRISPR bacterial immune system and has been repurposed for targeted gene editing, providing major advances in efficiency, speed and scale of genome engineering[12,52,55,57]. The endonuclease Cas9 has two nuclease domains, RuvC and HNH, and is directed to specific DNA sequences by pairing of a chimeric single guide RNA (sgRNA) with target DNA (see the figure)[13,29]. This requires the presence of a 5′ protospacer adjacent motif (PAM) in the DNA, which in *Streptococcus pyogenes* is usually NGG.

Genome engineering with Cas9 exploits DNA double-strand break (DSB) repair pathways to create mutations at specific genomic locations. In mammalian cells, DSBs are usually repaired either by homology-directed repair (HDR), for which an exogenous template is provided for the conversion of the endogenous sequence following the formation of DSBs by Cas9, or by non-homologous end joining (NHEJ), which is an error-prone mechanism that generates random insertion and deletion mutations at the targeted location. The continuous expression of the Cas9–sgRNA complex leads to near-complete allelic modification in a short time frame[58]. However, the use of Cas9 nuclease is potentially associated with the formation of off-target DSBs, which may cause undesired effects (BOX 3).

The enzymatic activity of Cas9 can be abolished by mutations in the RuvC and HNH domains, generating nuclease-deactivated Cas9 (dCas9), which can still be specifically targeted to DNA[89]. dCas9 can modulate transcription when fused to an effector domain (for example, a transcription activator or repressor) and expressed in concert with a sgRNA targeted to a regulatory element of interest[16,29]. An advantage of using dCas9 effectors is that their effect is reversible; moreover, dCas9 fusions can be multiplexed and set to perform simultaneous transcription activation and repression of different loci[71]. Potential disadvantages of dCas9 systems include disruption of native chromatin architecture, steric hindrance with transcription factors[24], and blocking of RNA polymerases[29]. In addition, the efficiency of activation or repression can fluctuate dramatically[32,34]. This can be explained by features of the target region, such as chromatin structure, and by the availability of additional cofactors[32]. Figure adapted from REF. 90, Nature Publishing Group.

Genetic manipulation of enhancers

The genomic features that define a regulatory element are poorly understood. For example, some enhancers consist of a single unit, whereas others — known as super-enhancers — are composed of multiple clusters of enhancers[39]. Genetic perturbation is a powerful approach to draw causal links between genetic information and cellular functions. The advent of the CRISPR–Cas9 system facilitated genome editing and, recently, it has been applied to dissect the DNA sequences required for appropriate activity of enhancer elements in their native context.

The Cas9 nuclease can be directed by one sgRNA to induce a double-strand break (DSB) at a specific genomic region. These DSBs are generally repaired by the error-prone repair pathway non-homologous end joining (NHEJ), resulting in the formation of insertion and deletion (indel) mutations (usually smaller than 10 bp). The initial application of CRISPR–Cas9 for genome editing was centred on protein-coding genes, but it was soon adopted to target non-coding regulatory elements. Recently, it was found that CRISPR–Cas9 targeting by one sgRNA can produce comparable genetic effects to the deletion of a whole enhancer unit[18] and cause phenotypes that are robust enough for screening[19] (see below). The strong effects produced by one sgRNA go against the well-established idea that regulatory elements are robust and redundant. Indeed, functionally important sequences within the enhancer are known to be highly sensitive as well: even single-nucleotide alterations in these regulatory sequences can have substantial effects on gene expression and cause pathological conditions[3,4,6–8]. In line with this, it was recently proposed that small mutations can activate oncogenes in cancer cells by creating a new super-enhancer[5]. In this study, a single Cas9–sgRNA complex was used to delete the mutant sequence, resulting in the collapse of enhancer activity and decreased expression of the *TAL1* (T-cell acute lymphocytic leukemia protein 1) proto-oncogene.

We note several limitations when using a single Cas9–sgRNA complex to target enhancer elements. First, one sgRNA might not fully disrupt enhancer activity owing to the small size of indels generated. The use of paired sgRNAs to delete genomic elements can address this issue (see bellow). Second, when working with a population of cells, the effect of a sgRNA might be diluted, as targeting by Cas9 results in different (heterozygous and homozygous) lesions in different cells. The isolation of individual mutant clones is a laborious process, but it can potentially solve this problem. Finally, some sgRNAs may suffer from intrinsic low DNA-editing efficiency (BOX 3).

A regulatory element can be fully inactivated by Cas9 nuclease through the generation of a specific deletion. This can be easily achieved by targeting two sgRNAs to the borders of the target region. An example of this strategy comes from a report in which CRISPR–Cas9 was used to excise an enhancer that regulates *GATA2* (GATA binding protein 2) expression[40]. The authors showed that this regulatory element is frequently translocated in leukaemia cells, leading to haploinsufficiency of *GATA2* and activation of *EVI1* (ecotropic viral integration site 1) oncogene. Cas9 nuclease can also be used to generate monoallelic deletions of enhancer elements. When targeting essential regulatory elements, a monoallelic deletion is performed to circumvent the cellular lethality associated with biallelic deletions. The combination of this targeted approach with gene expression profiling identified a distal enhancer cluster that is required for *Sox2* (SRY (sex determining region Y)-box 2) transcription[41,42] and ES cell differentiation[41]. Interestingly, deletion of this cluster did not affect other nearby genes, indicating that the regulation of gene expression by enhancers is highly specific[41,42]. The high precision of Cas9-mediated cleavage allows the deletion of individual constituents of composite enhancers[18,33,43]. Recent evidence suggests that deletion of a single element can have comparable effects to the deletion of the entire composite enhancer[18]. More studies are needed in order to understand whether

Box 3 | **Targeting specificity of CRISPR–Cas9 systems**

Targeting accuracy of CRISPR–Cas9 systems is defined as the ratio between generation of DNA double-strand breaks (DSBs) at the intended locations (on-target effects) and at unintended locations (off-target effects) and is assessed for each single guide RNA (sgRNA) in terms of their consistency in targeting the same locus[54,56–58]. Nuclease-deactivated Cas9 (dCas9)-based screening (CRISPR-mediated gene activation (CRISPRa) and CRISPR interference (CRISPRi)) can modulate gene expression, and recent dCas9-based tiling screens improved their on-target effects and sgRNA consistency (that is, different sgRNAs efficiently targeting same region)[17,23]. The use of dCas9 for functional enhancer screens has not yet been evaluated, but the enzymatically active CRISPR–Cas9 system was recently demonstrated to be a powerful enhancer screening tool[18,19,46]. Nevertheless, sgRNAs display large variations in efficiency (a result of chromatin accessibility and the composition of the targeted sequence[54,91,92]), which may influence screening results.

The continuous expression of Cas9 leads to near-complete allelic modification owing to the irreversible modification of target DNA[54,56,57]. However, not every mutation will abolish gene function as DSB repair by non-homologous end joining generates different mutations in different cells. Moreover, the size of insertion and deletion (indel) mutations varies substantially between the targeted loci[54,56,57,93]. Generating a DSB in a constitutively spliced exon usually results in the formation of either a premature stop codon or frameshift mutations[44,92]. By contrast, the critical regulatory elements of enhancers are poorly defined, and therefore it is challenging to target them with one sgRNA. Cas9-induced mutations in regulatory elements can result in more variable effects on gene expression compared to mutations in protein-coding regions[45,46]. This is probably caused by a wide spectrum of mutations with different likelihood to disrupt critical DNA elements, such as transcription factor binding sites.

The specificity of CRISPR–Cas9 reagents is a major concern for their applicability in both research and clinical settings. Cas9 nucleases can tolerate up to five mismatches between the sgRNA and a genomic sequence[94,95]. Reassuringly, the majority of these off-target binding events do not lead to the generation of indels[91,96], suggesting that imperfect matching of the sgRNA to the target sequence is insufficient for DNA cleavage[97]. dCas9-based tools were shown to be sensitive to even fewer mismatches[17], and global gene expression analysis revealed highly specific activity of both transcriptional and epigenetic modulators[16,17,28]. Nonetheless, CRISPR reagents with improved specificity were developed, such as paired Cas9 nucleases that nick DNA at two adjacent sequences[53,98], truncated sgRNAs[99], and modified Cas9 proteins with high precision[100,101]. The D10A mutant of Cas9 nicks one strand of DNA, and directing paired Cas9(D10A) nucleases to a specific location results in fewer off-target effects compared with Cas9 (REFS 53,98). Truncated sgRNAs, which are designed according to the finding that PAM-distal regions of sgRNA usually have less effect on on-target cleavage efficiency, have reduced off-target effects (up to 5,000 times) while conserving on-target cleavage efficiency[99].

Increasing CRISPR–Cas9 fidelity may also reduce false-negative rates in a screen (true hits that were missed in the screen). To our knowledge, false-negative rates have not been reported so far for CRISPR–Cas9 enhancer screens. It is possible that the effects of sgRNAs with low activity are missed, resulting in higher false-negative rates; if this is the case improving target specificity and sgRNA consistency should improve genetic screening approaches by reducing false-positive as well as false-negative rates.

this is an exception or a rule. The use of paired sgRNAs enables control of the size of the deletion, in contrast with the random indels induced by using a single sgRNA. However, this method is far less efficient in inducing deletions[44]: the efficiency of targeted deletions by paired sgRNAs is approximately 25%, but it decreases substantially with increasing deletion sizes[44]. In addition, the introduction of two sgRNAs theoretically carries twice the risk of off-target effects (BOX 3) compared with one sgRNA.

These studies emphasize the importance of perturbing the endogenous state of enhancers to obtain greater insight into their biological function. The combination of CRISPR–Cas9-induced mutagenesis and genome-wide expression profiling allows the identification of enhancers and their target genes, which is a major advantage compared to previous technologies (BOX 1). This combination is suitable for matching an individual enhancer with its target gene, and cannot yet be used on a genome-wide scale. It is also worth noting that the function of some *cis* regulatory elements might be redundant (see below). This problem can be tackled by targeting Cas9 nuclease to multiple redundant enhancers and abolishing their activity in a combinatorial manner.

High-throughput screening of enhancers

The power of the CRISPR–Cas9 system relies on two main factors: the ease of RNA-mediated gene targeting and its efficient induction of mutations[12,14]. The capacity of targeting Cas9 to a specific locus by a short (20-nucleotide) sgRNA enables the rapid generation of sgRNA libraries for screening. The sgRNAs can be delivered by a lentiviral vector and stably integrated in the target cells, facilitating the screen readout using next-generation sequencing (FIG. 1). Recently, several research groups have taken advantage of these features to perform high-throughput functional screens of endogenous enhancers.

Dissecting enhancers using CRISPR–Cas9 tiling screens. Single-nucleotide changes can substantially affect gene regulation[44], and this finding led to the hypothesis that Cas9-mediated mutagenesis could uncover elements required for the function of enhancers. The basic strategy is to tile hundreds of sgRNAs across a putative enhancer to disrupt nearly its entire sequence (FIG. 2). The screen is then used to infer which enhancer sequences are potentially important for its function. However, the full spectrum of the driver mutations and potential regulatory DNA elements can only be examined at the validation stage using individual sgRNA transductions, as sgRNA enrichment scores might differ from expectations[19]. Importantly, the resolution of such tiling screens approaches saturation mutagenesis *in situ*, with average cleavage frequency ranging from 4 to 10 bp[18,19]. The application of this method enables the study of the regulation of single genes by multiple enhancer elements[18,19,45,46].

Tiling screens were used successfully to dissect the human and mouse *BCL11A* (B-cell lymphoma/leukaemia 11A) composite enhancer[18]. Interestingly, the critical components of the composite enhancer were found to be different between human and mouse[18], despite sharing sequence homology and chromatin signatures. As BCL11A regulates the switching from fetal to adult haemoglobin gene expression by repressing the expression levels of fetal heamoglobin, the results from this study suggest that genetic perturbation of the *BCL11A* enhancer might be a therapeutic option for β-haemoglobin disorders. Similarly, a CRISPR–Cas9 tiling screen to identify critical sequences that contribute to the function of a distal enhancer of the p53 target gene *CDKN1A*[19] revealed that additional domains, besides the p53 binding site, are required for the activity of the enhancer and activation of *CDKN1A* expression upon oncogene-induced senescence. High-throughput CRISPR–Cas9 mutagenesis was also employed to systematically interrogate putative regulatory elements of *POU5F1* in human ES cells[46]. In this case, a POU5F1–GFP reporter was used to determine the function of the candidate enhancer regions. The authors identified classical regulatory elements, but also a class of non-canonical enhancers, termed temporarily phenotypic (Temp) enhancers. Although they harbour common enhancer features, loss of these elements caused temporary and reversible transcriptional impairment owing to disruption of the local chromatin structure. Further work is required in order to understand the biological function of Temp enhancers and how they are regulated.

Another high-throughput mapping of regulatory sequences is multiplexed editing regulatory assay (MERA)[45]. It is based on the integration of a pooled sgRNA library into a specific locus by homologous recombination and relies on the expression of a GFP-tagged gene locus to determine the effect of the mutations[45]. MERA led to the identification of expected promoters, enhancers and transcription factor binding sites, but also what might be a new class of elements, designated as unmarked regulatory elements (UREs). UREs were shown to regulate endogenous gene expression although they do not contain any known epigenetic markers associated with regulatory elements. These results expose the shortcomings of correlative annotations to predict regulatory sequences and highlight the importance of direct perturbation to definitively characterize genetic elements.

Considering these successful experiments, high-throughput CRISPR–Cas9 tiling screens should be readily implemented to interrogate at near-nucleotide resolution the function of numerous non-coding regulatory elements. The Cas9 protein scans for a short DNA sequence known as protospacer adjacent motif (PAM; in *Streptococcus pyogenes* it

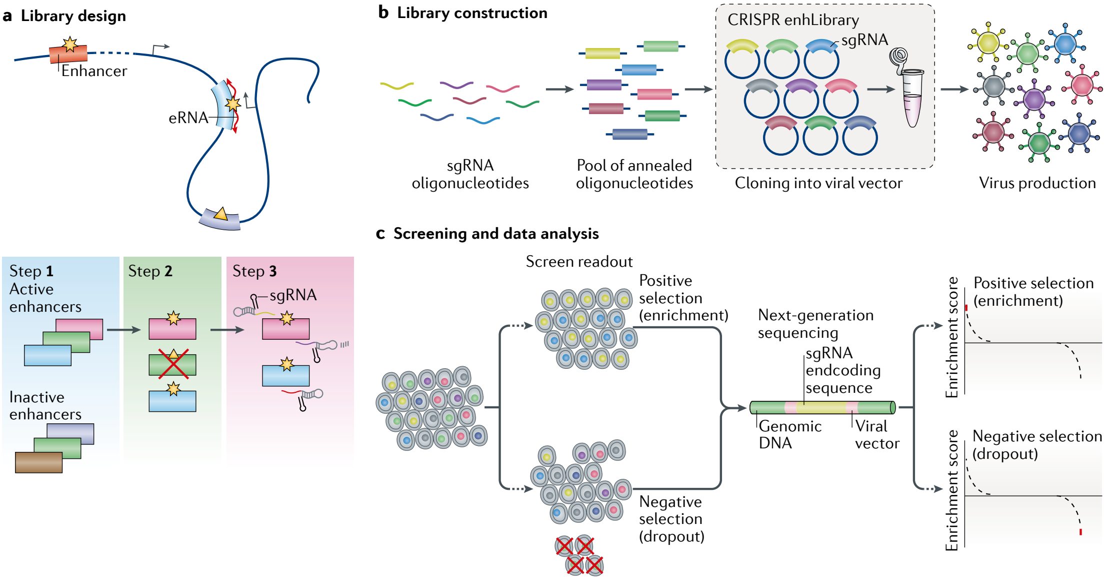


Figure 1 | **Functional genetic screens of active enhancers using the CRISPR–Cas system. a** | Bioinformatic design of a single guide RNA (sgRNA) library for global targeting of putative enhancer elements consists of three steps[19]. The first step is identification of the enhancers active in the specific experimental condition. The chromatin at active enhancers is typically marked with specific epigenetic modifications (high level of H3K4me1 (histone H3 Lys4 monomethylation) and H3K27ac (histone H3 Lys27 acetylation), and low level of H3K4me3 (histone H3 Lys4 trimethylation)) and occupied by RNA polymerase II. Thus, publicly available ChIP–seq (chromatin-immunoprecipitation followed by sequencing) data sets of chromatin modifications and of RNA polymerase II occupancy can be used to identify putative enhancer regions. Furthermore, enhancer RNA (eRNA) expression correlates well with enhancer activity and allows active regulatory elements to be distinguished; eRNA expression data sets are also publicly available. The second step is identification of active enhancer regions harbouring transcription factor binding sites (transcription factor binding sites (for example, p53 and oestrogen receptor-α (ERα)) are shown by yellow star and triangle symbols). Enhancer regulation depends on binding of specific transcription factor combinations, and each enhancer is responsive to a different set of transcription factors. Therefore, the intersection of active enhancer regions (as defined in step 1) with transcription factor activity relevant to the specific experimental condition (marked with a star) results in a final list of enhancers to be targeted by CRISPR–Cas9. The third step is identification of suitable Cas9 cleavage sites within the transcription factor binding sites, according to the availability of protospacer adjacent motif (PAM) sequences. It is essential to target precisely the transcription factor binding site within the enhancer region; yet, the availability of such targetable sites varies according the availability of PAMs in the transcription factor consensus sequence. For instance, the consensus motif of p53 has high NGG PAM content compared to ERα (resulting in ~90% versus ~60% target availability, respectively), and some regions can be targeted with more than one sgRNA. **b** | The sgRNA oligonucleotides targeting the suitable Cas9 cleavage sites are synthesized, annealed with 3′ and 5′ cloning primers, pooled and cloned into viral constructs to produce an enhancer-targeting sgRNA expression library (CRISPR enhLibrary), from which viruses are produced to confer stable expression of sgRNAs in cells. **c** | Virus transduction of cells should ideally be performed such that each cell expresses only one sgRNA, but that all the sgRNAs are expressed in the transduced cell population, to maintain the complexity of the library. The transduced cells are subjected to a proliferation-based screening selection to identify sgRNAs that confer cell growth advantage or disadvantage according to the designed assay, and next-generation sequencing is used to assess which sgRNAs were enriched or depleted (shown in red) in the selected cell population.

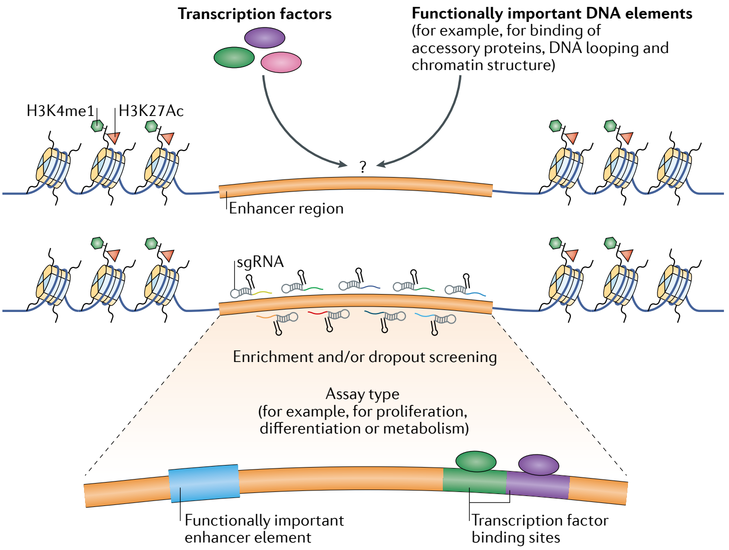

Transcription factors
Functionally important DNA elements
(for example, for binding of accessory proteins, DNA looping and chromatin structure)
H3K4me1
H3K27Ac
?
Enhancer region
sgRNA
Enrichment and/or dropout screening
Assay type
(for example, for proliferation, differentiation or metabolism)
Functionally important enhancer element
Transcription factor binding sites

Figure 2 | **High-resolution dissection of active enhancer sequences (marked by H3K4me1 and H3K27ac) by CRISPR–Cas9 tiling screens to uncover novel regulatory elements.** To gain detailed understanding of functionally important DNA elements (represented by the question mark) within enhancers, such as those important for binding of transcription factors or mediating DNA looping, all protospacer adjacent motif sites within a selected enhancer are identified by bioinformatics, and single guide RNAs (sgRNAs) densely covering (tiling) the region are produced and used for screening by an assay of interest. Following screening, the abundance of sgRNAs in the final cell population can be used to infer key elements that are required for the function of the enhancer, such as novel transcription binding sites. H3K4me1, histone H3 Lys4 monomethylation; H3K27ac, histone H3 Lys27 acetylation.

is usually NGG) at the 3′ end of the target sequence to achieve a successful binding to the target region, and generates DSBs 3 bp upstream of the PAM. Therefore, the use of Cas9 orthologues and variants, which have different PAM requirements, could increase the resolution of CRISPR–Cas9 *in situ* mutagenesis[47–50]. Alternatively, activating homology-directed repair (HDR) concomitantly with CRISPR–Cas9 could allow DNA editing at nucleotide resolution. The coupling of CRISPR–Cas9 mutagenesis and multiplex HDR has already enabled saturation editing of protein-coding genes[51], suggesting that it can also be applied to dissect the function of regulatory elements with high resolution. However, we note that HDR in conjugation with CRISPR–Cas9 has low DNA-editing efficiency, that it is time consuming and might not be possible for every cell type.

In high-throughput screens of protein-coding genes, false identification of screen hits (false positives) can be avoided by designing a library containing multiple unique sgRNAs targeting the same gene. Only screen hits with distinct sgRNAs yielding the same phenotype are considered. Also, rational design of sgRNAs can minimize false positives owing to off-target effects[52,53]. It was shown that the PAM-proximal 10–12 bp of sgRNA sequence is the main determinant of specificity; thus, restraining the similarity between the target and off-target sites and allowing no more than three mismatches can reduce the off-target effects. However, when targeting regulatory elements with CRISPR–Cas9, the availability of PAM sequences is a limiting factor in the design of multiple unique sgRNAs targeting the same region. In this case, the performance of single sgRNAs should be evaluated in several independent experiments to identify a consistent effect. Finally, the inclusion of many control sgRNAs will facilitate the distinction between true- and false-positive screen results.

Genetic screens of enhancers. In recent years, CRISPR–Cas9-based genetic screens have been applied successfully to identify protein-coding genes involved in fundamental cellular processes[54,55] and resistance to drugs and toxins[41,56,57]. Likewise, high-throughput CRISPR–Cas9 strategies are suitable for investigating the role of multiple enhancers in the regulation of multiple genes and their contribution to a specific phenotype for large-scale genetic screening of regulatory elements[19].

Indeed, CRISPR–Cas9 was used in both positive and negative selection screens to identify functional enhancers[19] by inactivating putative enhancers through the disruption of transcription factor binding sites and screening for phenotypic changes (FIG. 1). As a result, several novel enhancers were identified, which have key roles in p53-dependent oncogene-induced senescence and oestrogen receptor 1 (ESR1)-mediated cell proliferation[19]. This screening method was found to be efficient and robust, generating high levels of mutagenesis, sgRNA consistency (that is, different sgRNAs with similar efficiency in targeting the same region) and strong phenotypic effects. In particular, the performance of Cas9 in the negative selection setting, as determined by high validation rates, was impressive[19], as this type of screen requires higher sensitivity to detect changes in the representation of individual sgRNAs[58].

Studies discussed above[18,19,45,46] have produced some common results, despite being performed in completely distinct biological settings. First, both enhancers and super-enhancers can be susceptible to small mutations generated by single DSBs, which enabled these high-throughput functional screens[18,19,45,46]. However, we cannot rule out that some enhancers are robust and not affected by small indels. The development of methods to generate pooled libraries of paired sgRNAs can potentially address this limitation[59]. Second, the vast majority of sgRNAs were not enriched or depleted, whereas the ones that were, colocalized to discrete genomic regions[18,19]. This suggests that enhancer elements are composed of many redundant and only a few critical sequences. Finally, the pattern of mutations generated by a sgRNA can be used for motif discovery without a priori knowledge[18,19,45]. This can be done by mutagenizing enhancers in a population of cells using multiple, unique sgRNAs, applying selective conditions and comparing the initial and final abundance of mutations along the enhancer regions. In principle, the relative abundance of mutations at specific sequences should reflect their necessity for the function of an enhancer element. Altogether, these studies establish CRISPR–Cas9 as a powerful tool to dissect the functions of the non-coding genome in an unbiased fashion[18,19,45,46].

A caveat of the sgRNA libraries used to date is that they target a limited set of enhancers and therefore cannot be used for unbiased screening in any biological setting. We envision that, in the near future, it will be possible to interrogate the function of all genomic regulatory elements using CRISPR–Cas9 tools. This will be a monumental task, given that the estimated number of enhancers in the human genome is estimated to be approximately one million (REFS 10,60). The identification of functional enhancers by CRISPR–Cas9 on a genome-wide scale will require further technological advances to improve the coverage of libraries and speed of screening.

Future perspective

Since it was first used in genome engineering in 2013 (REFS 12–15), the CRISPR–Cas9 system and its applications have continued to improve at an unprecedented rate. Here, we pinpoint future directions that can be applied to study transcriptional regulatory regions with this system.

Light-inducible systems. The development of light-inducible DNA binding proteins represents a breakthrough for genome engineering. Initially, this technology was applied to transcription activator-like effector proteins (TALEs)[61] but it has recently been introduced to CRISPR–Cas9 systems. These tools are incredibly versatile as they can be activated within minutes and do not require additional chemical cofactors. Light-inducible CRISPR–Cas9 systems have been used to control genome editing[62] and modulate gene expression[63,64] in a reversible manner. In principle, this technology can be adapted to modulate the activity of endogenous enhancers in space and time with very good resolution.

Modelling human diseases. The CRISPR–Cas9 system simplified and accelerated the modelling of complex human diseases. For example, paired sgRNAs were used to engineer a chromosomal inversion and generate mouse models of human non-small cell lung cancers[65]. This approach can also be used to investigate human diseases associated with repositioning of enhancers by chromosomal rearrangements[40]. In recent years, it has become clear that the vast majority of disease-associated mutations and single-nucleotide polymorphisms are present in non-coding regions of the genome[9,66]. The clinical significance of these genetic alterations is unknown as it is not clear whether they affect gene expression. The combination of CRISPR–Cas9-mediated mutagenesis and multiplex HDR promises to facilitate the interpretation of sequence variants of uncertain significance reported by clinical sequencing[51].

Manipulation of high-order chromatin organization. Transcriptional regulation can also be achieved by customized editing of the genome architecture. It was recently demonstrated that disruption of topological boundaries by Cas9 nuclease resulted in loss of facultative heterochromatin and activation of gene expression[67]. Mutations affecting topological boundaries were found in different types of cancer, and this was recently proposed to be a mechanism of activation of proto-oncogenes by enhancer elements[68]. Forced chromatin looping using a synthetic zinc finger–LDB1 fusion protein resulted in the activation of developmentally silenced genes in mouse cells[69,70]. This suggests that programmable DNA looping could be a novel therapeutic approach for human diseases. Although there are no CRISPR–Cas-based tools available to induce changes in chromatin conformation, dCas9 has the potential to bring together two or more genomic loci through the use of sgRNAs that scaffold multiple domains, which are capable of recruiting Cas9 and other sgRNAs[71]. The spread of this strategy to CRISPR–Cas9 systems could help to characterize the role of three-dimensional genome conformations in the function of regulatory elements and gene expression.

Concluding remarks. The future application of CRISPR–Cas9 systems in genetic screens of enhancers will certainly accelerate the identification of regulatory networks by connecting non-coding elements to their target genes and to a specific phenotype. We believe that the most comprehensive definition of the function of regulatory elements will arise from combining different strategies, including disruption of the DNA sequence and modulation of its activity. However, the majority of CRISPR–Cas9 tools have only been applied to study a limited number of endogenous enhancers to date. It is paramount to evaluate these tools across many regulatory elements and in different biological settings to fully appreciate their potential. Together, we are privileged to live in the 'CRISPR–Cas9 era', a time in which our imagination seems to be the only limiting factor in discovering the secrets of the human genome.

Rui Lopes, Gozde Korkmaz and Reuven Agami are at the Division of Biological Stress Response, The Netherlands Cancer Institute, Plesmanlaan 121, 1066CX Amsterdam, The Netherlands.

Reuven Agami is also at the Department of Genetics, Erasmus University Medical Center, Wytemaweg 80, 3015 CN Rotterdam, The Netherlands.

R.L. and G.K. contributed equally to this work.

Correspondence to G.K. and R.A.
g.korkmaz@nki.nl; r.agami@nki.nl

doi:10.1038/nrm.2016.79
Published online 6 July 2016;
corrected online 22 July 2016

1. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
2. Bulger, M. & Groudine, M. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev. Biol.* **339**, 250–257 (2010).
3. Bauer, D. E. *et al.* An erythroid enhancer of *BCL11A* subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253–257 (2013).
4. Harismendy, O. *et al.* 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature* **470**, 264–268 (2011).
5. Mansour, M. R. *et al.* An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377 (2014).
6. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
7. Musunuru, K. *et al.* From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
8. Oldridge, D. A. *et al.* Genetic predisposition to neuroblastoma mediated by a *LMO1* super-enhancer polymorphism. *Nature* **528**, 418–421 (2015).
9. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
10. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
11. Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).
12. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
13. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *eLife* **2**, e00471 (2013).
14. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
15. Cho, S. W., Kim, S., Kim, J. M. & Kim, J. S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.* **31**, 230–232 (2013).
16. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).
17. Gilbert, L. A. *et al.* Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* **159**, 647–661 (2014).
18. Canver, M. C. *et al.* *BCL11A* enhancer dissection by Cas9-mediated *in situ* saturating mutagenesis. *Nature* **527**, 192–197 (2015).
19. Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat. Biotechnol.* **34**, 192–198 (2016).
20. Maeder, M. L. *et al.* Targeted DNA demethylation and activation of endogenous genes using programmable TALE–TET1 fusion proteins. *Nat. Biotechnol.* **31**, 1137–1142 (2013).
21. Perez-Pinera, P. *et al.* RNA-guided gene activation by CRISPR–Cas9-based transcription factors. *Nat. Methods* **10**, 973–976 (2013).
22. Maeder, M. L. *et al.* CRISPR RNA-guided activation of endogenous human genes. *Nat. Methods* **10**, 977–979 (2013).
23. Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583–588 (2015).
24. Tanenbaum, M. E., Gilbert, L. A., Qi, L. S., Weissman, J. S. & Vale, R. D. A protein-tagging system for signal amplification in gene expression and fluorescence imaging. *Cell* **159**, 635–646 (2014).

25. Choy, B. & Green, M. R. Eukaryotic activators function during multiple steps of preinitiation complex assembly. *Nature* **366**, 531–536 (1993).
26. Memedula, S. & Belmont, A. S. Sequential recruitment of HAT and SWI/SNF components to condensed chromatin by VP16. *Curr. Biol.* **13**, 241–246 (2003).
27. Gao, X. *et al.* Comparison of TALE designer transcription factors and the CRISPR/dCas9 in regulation of gene expression by targeting enhancers. *Nucleic Acids Res.* **42**, e155 (2014).
28. Hilton, I. B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* **33**, 510–517 (2015).
29. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).
30. Chen, B. *et al.* Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479–1491 (2013).
31. Crocker, J. & Stern, D. L. TALE-mediated modulation of≈transcriptional enhancers *in vivo*. *Nat. Methods* **10**, 762–767 (2013).
32. Thakore, P. I. *et al.* Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods* **12**, 1143–1149 (2015).
33. Zhang, X. *et al.* Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat. Genet.* **48**, 176–182 (2016).
34. Kearns, N. A. *et al.* Functional annotation of native enhancers with a Cas9–histone demethylase fusion. *Nat. Methods* **12**, 401–403 (2015).
35. Groner, A. C. *et al.* KRAB–zinc finger proteins and KAP1 can mediate long-range transcriptional repression through heterochromatin spreading. *PLoS Genet.* **6**, e1000869 (2010).
36. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* **16**, 144–154 (2015).
37. Shlyueva, D. *et al.* Hormone-responsive enhancer-activity maps reveal predictive motifs, indirect repression, and targeting of closed chromatin. *Mol. Cell* **54**, 180–192 (2014).
38. Shi, Y. *et al.* Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell* **119**, 941–953 (2004).
39. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
40. Groschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant *EVI1* and *GATA2* deregulation in leukemia. *Cell* **157**, 369–381 (2014).
41. Zhou, H. Y. *et al.* A *Sox2* distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes Dev.* **28**, 2699–2711 (2014).
42. Li, Y. *et al.* CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS ONE* **9**, e114485 (2014).
43. Hnisz, D. *et al.* Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol. Cell* **58**, 362–370 (2015).
44. Canver, M. C. *et al.* Characterization of genomic deletion efficiency mediated by clustered regularly interspaced palindromic repeats (CRISPR)/Cas9 nuclease system in mammalian cells. *J. Biol. Chem.* **289**, 21312–21324 (2014).
45. Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* **34**, 167–174 (2016).
46. Diao, Y. *et al.* A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res.* **26**, 397–405 (2016).
47. Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Methods* **10**, 1116–1121 (2013).
48. Kleinstiver, B. P. *et al.* Broadening the targeting range of *Staphylococcus aureus* CRISPR-Cas9 by modifying PAM recognition. *Nat. Biotechnol.* **33**, 1293–1298 (2015).
49. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).
50. Ran, F. A. *et al.* *In vivo* genome editing using *Staphylococcus aureus* Cas9. *Nature* **520**, 186–191 (2015).
51. Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513**, 120–123 (2014).
52. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
53. Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31**, 833–838 (2013).
54. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).
55. Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910–918 (2013).
56. Koike-Yusa, H., Li, Y., Tan, E. P., Velasco-Herrera Mdel, C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.* **32**, 267–273 (2014).
57. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
58. Shalem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR–Cas9. *Nat. Rev. Genet.* **16**, 299–311 (2015).
59. Vidigal, J. A. & Ventura, A. Rapid and efficient one-step generation of paired gRNA CRISPR-*Cas9* libraries. *Nat. Commun.* **6**, 8083 (2015).
60. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
61. Konermann, S. *et al.* Optical control of mammalian endogenous transcription and epigenetic states. *Nature* **500**, 472–476 (2013).
62. Nihongaki, Y., Kawano, F., Nakajima, T. & Sato, M. Photoactivatable CRISPR-Cas9 for optogenetic genome editing. *Nat. Biotechnol.* **33**, 755–760 (2015).
63. Nihongaki, Y., Yamamoto, S., Kawano, F., Suzuki, H. & Sato, M. CRISPR-Cas9-based photoactivatable transcription system. *Chem. Biol.* **22**, 169–174 (2015).
64. Polstein, L. R. & Gersbach, C. A. A light-inducible CRISPR-Cas9 system for control of endogenous gene activation. *Nat. Chem. Biol.* **11**, 198–200 (2015).
65. Maddalo, D. *et al.* *In vivo* engineering of oncogenic chromosomal rearrangements with the CRISPR/Cas9 system. *Nature* **516**, 423–427 (2014).
66. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
67. Narendra, V. *et al.* CTCF establishes discrete functional chromatin domains at the *Hox* clusters during differentiation. *Science* **347**, 1017–1021 (2015).
68. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).
69. Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233–1244 (2012).
70. Deng, W. *et al.* Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell* **158**, 849–860 (2014).
71. Zalatan, J. G. *et al.* Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. *Cell* **160**, 339–350 (2015).
72. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
73. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
74. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
75. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
76. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
77. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
78. Fullwood, M. J. *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
79. Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
80. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
81. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
82. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nat. Biotechnol.* **30**, 265–270 (2012).
83. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
84. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol. Cell* **49**, 825–837 (2013).
85. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
86. Rivera, C. M. & Ren, B. Mapping human epigenomes. *Cell* **155**, 39–55 (2013).
87. Hah, N., Murakami, S., Nagari, A., Danko, C. G. & Kraus, W. L. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res.* **23**, 1210–1223 (2013).
88. Arner, E. *et al.* Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**, 1010–1014 (2015).
89. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
90. 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).
91. Wu, X. *et al.* Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32**, 670–676 (2014).
92. Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR–Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
93. Bae, S., Kweon, J., Kim, H. S. & Kim, J. S. Microhomology-based choice of Cas9 nuclease target sites. *Nat. Methods* **11**, 705–706 (2014).
94. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).
95. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).
96. Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.* **32**, 677–683 (2014).
97. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).
98. Ran, F. A. *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380–1389 (2013).
99. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32**, 279–284 (2014).
100. Kleinstiver, B. P. *et al.* High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
101. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).

Acknowledgements

The authors apologize to those whose work they were not able to mention or cite owing to space limitations. R.L. is a fellow of the Fundação para a Ciência e Tecnologia, Portugal (SFRH/BD/74476/2010; POPH/FSE). This work was supported by funds of enhReg ERC-AdV (ERC-2012-ADG -N°322493) and NWO (NGI 93512001/2012) to R.A.

Competing interests statement

The authors declare no competing interests.

Chapter 3

GRO-seq, a tool for identification of transcripts regulating gene expression

Adapted from Methods Molecular Biology. 2017;1543:45-55.

GRO-seq, A Tool for Identification of Transcripts Regulating Gene Expression

Rui Lopes*, Reuven Agami, and Gozde Korkmaz*

Abstract

The advent of next-generation sequencing (NGS) technologies has revolutionized the way we do research on gene expression. High-throughput transcriptomics became possible with the development of microarray technology, but its widespread application only occurred after the emergence of massive parallel sequencing. Especially, RNA sequencing (RNA-seq) has greatly increased our knowledge about the genome and led to the identification and annotation of novel classes of RNAs in different species. However, RNA-seq measures the steady-state level of a given RNA, which is the equilibrium between transcription, processing, and degradation. In recent years, a number of dedicated RNA-seq technologies were developed to measure specifically transcription events. Global run-on sequencing (GRO-seq) is the most widely used method to measure nascent RNA, and in recent years, it has been applied successfully to study the function and mechanism of action of noncoding RNAs. Here, we describe a detailed protocol of GRO-seq that can be readily applied to investigate different aspects of RNA biology in human cells.

Key words Sequencing, GRO-seq, Nascent transcription, Noncoding RNA, Enhancer, Running Head: GRO-seq

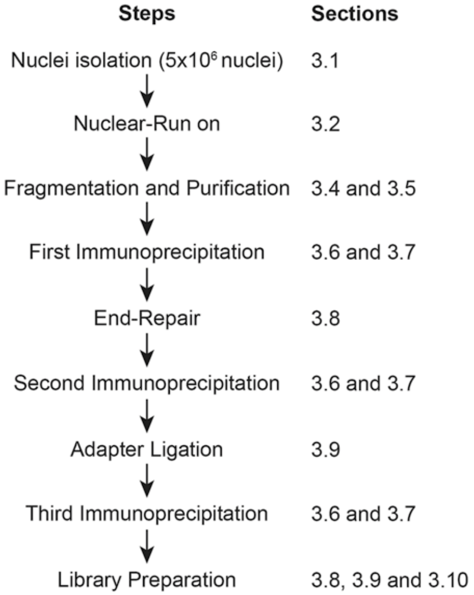
1 Introduction

RNA analysis was once limited to study individual molecules by Northern blot or real-time quantitative PCR. In the past decade, we witnessed a revolution in the RNA world caused by the rise of NGS. In particular, RNA-seq contributed decisively to make global gene expression analysis as a routine practice in many labs [1]. This technique does not only allow to qualitatively and quantitatively investigate messenger RNAs (mRNAs), but also novel noncoding RNA species (e.g., microRNAs, small interfering RNAs, and long noncoding RNAs) that are recognized nowadays as important players in the biology of the cell [2].

*These authors contributed equally to this work.

The cellular levels of a given RNA molecule are determined by the interplay of transcription, processing, and degradation. Consequently, both transcriptional and posttranscriptional changes affect RNA levels as measured by RNA-seq. In recent years, a number of RNA-seq-based technologies were developed that can measure specifically nascent RNA transcription from actively engaged polymerases. Among them, the most widely used are global run-on sequencing (GRO-seq) [3], native elongating transcript sequencing (NET-seq) [4], and precision nuclear run-on sequencing (PRO-seq) [5]. Here, we focus on GRO-seq, an assay that allows mapping and quantification of transcriptionally engaged RNA polymerases and provides a snapshot of genome-wide transcription. GRO-seq measurements are very sensitive and largely independent of RNA stability effects, making it particularly suitable to study noncoding RNA species that are lowly expressed and/or have high decay rate.

Recently, we and others have used GRO-seq to study different classes of noncoding RNAs such as promoter-associated RNAs [3], enhancer-associated RNAs [6], and long noncoding RNAs [7]. Below, we describe a detailed protocol of GRO-seq (Fig. 1), based on the original protocol from Core et al., Science (2008), which can be readily applied to study different aspects of RNA biology in human cells [8–10].



Steps	Sections
Nuclei isolation (5x10^6 nuclei)	3.1
Nuclear-Run on	3.2
Fragmentation and Purification	3.4 and 3.5
First Immunoprecipitation	3.6 and 3.7
End-Repair	3.8
Second Immunoprecipitation	3.6 and 3.7
Adapter Ligation	3.9
Third Immunoprecipitation	3.6 and 3.7
Library Preparation	3.8, 3.9 and 3.10

Fig. 1 Summarized scheme of the GRO-seq protocol steps (**left panel**) and corresponding sections (**right panel**) described in this article

2 Materials

All solutions for the protocol must be freshly prepared. In order to reduce the risk of RNase activity, it is necessary to use nuclease-free water for the preparation of all solutions, as well as the addition of RNase inhibitors to all RNA reactions.

2.1 Isolation of Nuclei

Buffers should be ice-cold during the protocol.

1. UltraPure DNase/RNase-Free Distilled Water.
2. PBS (DPBS, no calcium, no magnesium).
3. 1 MMgCl$_2$.
4. 0.5 M EDTA, pH 8.0.
5. UltraPure 1 M Tris–HCI, pH 8.0.
6. UltraPure 1 M Tris–HCI Buffer, pH 7.5.
7. Tris/HCl pH 7.8: mix 1:1 volume/volume ratio of Tris/HCl, pH 7.5 and Tris/HCl, pH 8.
8. RNase Inhibitor.
9. Swelling Buffer: 10 mM Tris/HCl pH 7.5, 2 mM MgCl$_2$ 3 mM CaCl$_2$. Add 2 U/ml RNase Inhibitor immediately before use.
10. Igepal CA-630.
11. Glycerol.
12. Lysis Buffer: 10 mM Tris–HCl pH 7.5, 2 mM MgCl$_{2,}$ 3 mM CaCl$_2$, 0.5 % (v/v) IGEPAL CA-630, 10 % (v/v) glycerol+ 2 U/ml RNase Inhibitor.
13. Freezing buffer: 40 % (v/v) glycerol, 5 mM MgCl$_2$, EDTA pH 8.0, 50 mM Tris/HCl pH 7.8, add 2 U/ml of SUPERase-In immediately before use.
14. Trypan blue solution.
15. Hemocytometer.
16. Liquid nitrogen.

2.2 NRO Reaction

1. 5-bromouridine 5′-triphosphate (BrUTP).
2. Sarkosyl 20 % (*N*-lauryl sarcosine sodium sulfate).
3. ATP Solution (10 mM).
4. CTP Solution (10 mM).
5. GTP Solution (10 mM).
6. KCl (2 M).

2.3 RNA Extraction

1. TRIzol LS (Invitrogen).
2. Chloroform.

3. GlycoBlue Coprecipitant (15 mg/mL) (Ambion).
4. 2-Propanol.

2.4 Fragmentation of NRO-RNA

1. RNA Fragmentation Reagents (Ambion).

2.5 Purification of Fragmented NRO-RNA Through p-30 RNase-free Spin Column

1. Micro Bio-Spin Columns with Bio-Gel P-30 in Tris Buffer (Bio-Rad).
2. TE buffer, pH 8.0.

2.6 Blocking the BrdU Beads for Immunoprecipitation

Add 2 μl of 20 U/μL RNase Inhibitor for each 10 ml of the buffers.

1. BrdU (IIB5) AC (Santa Cruz Biotech, catalogue number sc-32323 AC) (agarose conjugated BrdU antibody for IP studies).
2. UltraPure SSPE, 20×.
3. Tween 20.
4. UltraPure BSA (50 mg/ml).
5. PVP (Polyvinylpyrrolidone solution).
6. Binding Buffer: 0.25× SSPE, 0.001 M EDTA pH 8.0, 0.05 % Tween-20, 37.5 mM NaCl.
7. Blocking Buffer: 0.25× SSPE, 0.001 M EDTA pH 8.0, 0.05 % Tween-20, 37.5 mM NaCl, 0.1 % PVP (final concentration), 1 μg/ml BSA.

2.7 Immuno precipitation of NRO-RNA

Add 2 μl of 20 U/μL RNase Inhibitor for each 10 ml of the buffers.

1. PVP 40.000.
2. DTT.
3. 10 % SDS.
4. Low Salt Buffer: 0.25× SSPE, 0.001 M EDTA pH 8.0, 0.05 % (v/v) Tween-20.
5. High Salt Buffer: 0.25× SSPE, 0.001 M EDTA pH 8.0, 0.05 % (v/v) Tween-20, 100 mM NaCl.
6. TET Buffer: 1× TE buffer, 0.5 % (v/v) Tween-20.
7. Elution Buffer: 0.15 M NaCl, 0.05 M Tris pH 7.5, 0.001 M EDTA pH 8.0, 0.1 % SDS. Add 0.02 M DTT immediately before use.

2.8 End Repair of NRO-RNA (TAP/PNK Treatment)

1. Tobacco Acid Pyrophosphatase (TAP).
2. T4 Polynucleotide Kinase (PNK).

2.9 GRO-seq Library Preparation

1. TruSeq Small RNA Library Preparation Kit (Illumina).
2. Agencourt AMPure XP (Beckman Coulter).
3. RNA 6000 Ladder (Ambion).
4. Agilent DNA 1000 or Agilent High Sensitivity DNA chip.

3 Methods

3.1 Isolation of Nuclei

Perform all steps on ice or at +4 °C. Volumes below are described for 15 cm^2 plate.

1. Wash cells three times with ice-cold PBS.
2. Aspirate the PBS.
3. Add 10 ml of ice-cold Swelling Buffer to the plate, and incubate for 5 min on ice (*see* **Note 1**).
4. Scrape cells into the solution and transfer the solution into 15 ml tubes. Pellet cells by centrifuging at 400 × *g* (400 × *g*, GH3.8 rotor) for 10 min at +4 °C.
5. Aspirate supernatant (SN) and resuspend cells in 500 μl of Swelling Buffer/10 % glycerol solution containing 4 U/ml RNASe inhibitor by gentle pipetting.
6. Vortex cells slowly (≈800 rpm, so that liquid rises about 1–2 cm) and drop wise add 500 μl of Swelling Buffer/10 % glycerol/1 % IGEPAL CA-630 solution containing 4 U/ml RNASe inhibitor.
7. Incubate cells on ice for 5 min.
8. Bring volume to 10 ml with Lysis Buffer and centrifuge at 600 × *g* (600 × *g*) for 5 min at +4 °C.
9. Aspirate SN.
10. Wash nuclei with 10 ml of Lysis Buffer (*see* **Note 2**) and centrifuge at 600 × *g* (1550 rpm) for 5 min at +4 °C.
11. Aspirate the SN and resuspend the pellet in 1 ml of Freezing Buffer.
12. Mix 10 μl of Freezing buffer containing isolated nuclei with 190 μl Trypan Blue (2:5 diluted in freezing buffer, so that the nuclei will not swell) and count the cells by using hemocytometer.
13. Pellet nuclei at 900 × *g* (900 × *g*) for 6 min at +4 °C and aspirate the SN without disturbing the pellet.
14. Resuspend the pellet to have 5 × 10^6 nuclei per 100 μl with Freezing Buffer. Aliquots every 100 μl into individual 1.5 ml tube.
15. Snap-freeze nuclei in dry ice/methanol or liquid nitrogen or proceed directly to the nuclear run-on reaction (NRO-rxn) (*see* **Note 3**).

Table 1
Content of the NRO master mix (2×) for NRO Reaction

Reagent	Stock[M]	Final[mM]	500 µl
Tris–Cl	1	10	5
$MgCl_2$	1	5	2.5
DTT	0.1	1	5
KCl	2	300	75
ATP	0.01	0.5	25
GTP	0.01	0.5	25
CTP	0.0001	0.002	10
^{32}P-CTP			50
BrUTP	0.01	0.5	25
RNASe inhibitor	20 U/µl	0.4	10
2 % Sarkosyl	2	1	250
Water			17.5
			482.50
			500

3.2 NRO Reaction

1. Prepare the NRO master mix (2×) (Table 1) containing 1.165 µM final concentration of CTP (*see* **Note 4**).
2. Preheat the NRO-mix to +30 °C.
3. Mix 100 µl of pre-warmed NRO-mix with 100 µl nuclei (1:1 ratio) and place the reaction into the heat block (at +30 °C) to perform run-on for 5 min with shaking at the 2nd and 4th min (600–800 rpm) (*see* **Note 5**).
4. Immediately add 900–1000 µl of TRIzol LS and vortex solution for 30 s. This will stop the NRO-rxn (*see* **Note 6**).
5. Incubate for 5 min at RT.
6. It is advisable to keep the samples at −20 °C for short-term storage or at −80 °C for longer storage. OPTIONAL STOPPING POINT.

3.3 RNA Extraction

1. Add 240 µl of Chloroform to the 900–1000 µl of TRIzol LS and shake by hand for 15 s.
2. Incubate for 2–3 min at RT and centrifuge at 12,000 × *g* for 15 min at +4 °C.
3. Take out the colorless upper layer containing RNA (middle phase contains DNA, red phase contains proteins) and add 2 µl of glycogen and same volume of isopropanol as sample and inverse tube ten times (*see* **Note 7**).

4. Incubate for 10 min at RT and centrifuge at 14,000 × *g* for 15 min at +4 °C.
5. Remove supernatant and wash the pellet twice with 75 % EtOH (prepared with DNase, RNase free water), vortex, and centrifuge at 7500 × *g* for 5 min at +4 °C.
6. Remove supernatant, spin and remove the remaining of supernatant. Air-dry the pellet for 1 min at RT and be careful not to over-dry the pellet.
7. Suspend the pellet in 10 μl of DNase, RNase free water containing RNASe inhibitor enzyme (1 U/μl).
8. Incubate the solution for 5 min at 65 °C.

3.4 Fragmentation of NRO-RNA

1. Add 2 μl of fragmentation reagents up to 20 μl NRO-RNA solution and incubate for 10 min at +70 °C (*see* **Note 8**).
2. Add 2 μl of Stop solution and put on ice.

3.5 Purification of Fragmented NRO-RNA Through p-30 RNase-free Spin Column

1. In order to prepare column, invert it several times to resuspend the matrix. Spin at 1000 × *g* for 2 min to remove the flow through. Put column into a new 1.5 ml tube. Do not let the column dry out, prepare and use immediately.
2. Add 500 μl of TE buffer and centrifuge for 1 min at 1000 × *g*. Discard the flow through. Repeat this step for two to three times to change the buffer content.
3. Add sample (between 20 and 100 μl) into the column and centrifuge for 4 min at 1000 × *g*. Keep the flow through which contains fragmented NRO-RNA.
4. Check the radioactivity level with the Geiger counter. The flow through should have more counts than the matrix. If the matrix has more counts, it means that there is unincorporated ^{32}P-CTP. Thus, NRO-rxn was not successful.
5. Sample can be frozen at −80 °C. OPTIONAL STOPPING POINT.

3.6 Blocking the BrdU Beads for Immunoprecipitation

1. Equilibrate BrdU beads in binding buffer by washing them two times in 500 μl for 5 min with rotation (8 rpm). Spin down beads at 1000 × *g* for 1–2 min. Place on ice for 1 min before removing SN (*see* **Note 9**).
2. Block the beads in four to five times volume of blocking buffer for 1–2 h at RT. Add an extra 2 μl RNASe inhibitor for every ml of blocking buffer during this step.
3. Spin down the beads (1000 × *g* for 2 min) and discard the supernatant.
4. Wash beads twice in 500 μl binding buffer.

5. Spin down the beads (1000 × *g* for 2 min) and discard the supernatant.
6. Resuspend beads in 400 μl binding buffer/reaction.

3.7 Immuno precipitation (IP) of NRO-RNA (First IP)

1. Bring purified and fragmented NRO-RNA volume to 100 μl and add EDTA to reach a final concentration of 5 mM (add 1 μl of 0.5 M EDTA).
2. Heat the sample to +65–70 °C for 5 min and then place on ice for 2 min (*see* **Note 10**).
3. Add 400 μl of binding buffer and 50–60 μl of bead slurry into a 1.5 ml tube. Allow binding for +30–60 min (*see* **Note 11**).
4. Spin down the beads (1000 × *g* for 2 min) and discard the SN.
5. Wash one time in 500 μl of binding buffer for 5 min on rotating stand (8 rpm).
6. Spin down the beads (1000 × *g* for 2 min) and discard the SN.
7. Wash one time in 500 μl of low salt buffer for 5 min on rotating stand (8 rpm).
8. Spin down the beads (1000 × *g* for 2 min) and discard the SN.
9. Wash one time in 500 μl of high salt buffer for 3 min on rotating stand (8 rpm).
10. Spin down the beads (1000 × *g* for 2 min) and discard the SN.
11. Wash two times in 500 μl of TET buffer for 5 min on rotating stand (8 rpm).
12. Spin down the beads (1000 × *g* for 2 min) and discard the SN.
13. Elute two times with 125 μl and one time with 250 μl of elution buffer that is heated to +42 °C. Place the tube in a heat block (+42 °C) for 10 min with constant shaking at 500 rpm and every few minutes increase the shaking to 900 rpm.
14. Spin down the beads and transfer the solution (eluate) containing the NRO-RNA to a new tube.
15. Add 900–1000 μl of TRIzol LS and follow Subheading 3.3.

3.8 End Repair of NRO-RNA (TAP/PNK Treatment)

1. Heat RNA to +65–70 °C for 5 min, put on ice for 2 min.
2. Set up the following reaction in a 1.5 ml tube: 3 μl of 10× TAP buffer, 5 μl of nuclease-free water, 1 μl of RNASe inhibitor, 1.5 μl of TAP (10 U/ml).
3. Mix by pipetting and incubate the reaction at +37 °C for 1.5 h.
4. Add 1 μl of PNK, and 1 μl of 300 mM $MgCl_2$ to reach 10 mM final concentration, mix by pipetting, and incubate the reaction another 15 min.
5. Add 20 μl of PNK buffer, and 126 μl of nuclease-free water, 1 μl of RNAse inhibitor (20 U/μL), and another 1 μl of PNK.

6. Mix by pipetting and incubate the reaction for 15 min.
7. Add 20 μl of 10 mM ATP and another 1 μl of PNK, mix by pipetting, and incubate the reaction for 30 min.
8. Stop the reaction by adding 77 μl of nuclease-free water, 18 μl of 5 M NaCl, and 5 μl of 500 mM EDTA.
9. Add 900–1000 μl of TRIzol LS and follow Subheading 3.3.

3.9 Second IP, Adapter Ligation, and Third IP

1. Repeat Subheadings 3.6 and 3.7.
2. Follow the manufacturer's protocol for adapter ligation according to the desired sequencing platform. For example, TruSeq Small RNA kit should be used for Illumina sequencing system. As an alternative, custom protocols for adapter ligation have been previously described [11].
3. After 3′ and 5′ adapter ligation, repeat Subheadings 3.6 and 3.7.

3.10 Reverse Transcription (RT) and Cleaning-up with Magnetic Beads

1. Continue with the manufacturer's protocol for RT-PCR followed by Agencourt AMPure XP cleaning protocol. Use 1:2 (v/v) ratio for mixing sample and beads.
2. Elute in 25 μl for PCR Amplification.

3.11 PCR Amplification of the GRO-seq Library and Cleaning-up with Magnetic Beads

1. Continue with the manufacturer's protocol for PCR (max 12 cycles) followed by Agencourt AMPure XP cleaning protocol. Use 1:1 (v/v) ratio for mixing beads and sample.
2. Repeat magnetic beads cleaning step.
3. Elute in 15 μl.

3.12 Quantification of GRO-seq Library

1. Run 1 μl of the GRO-seq library on a Bioanalyzer using an Agilent DNA 1000 or Agilent High Sensitivity DNA chip. The fragments should be dispersed between 200 and 400 bp (Fig. 2).

3.13 Sequencing of GRO-seq Library and Data Analysis

1. Amplicons can be sequenced in a HiSeq 2500 (Illumina) platform using a standard 65 bp single-read reaction.
2. Sequenced reads can be aligned to the human genome (hg19) using bowtie2 [12] tool.
3. HOMER software can be applied to detect transcriptional units (TUs) [13]. The expression level of each TU can be calculated in each sample by using HTseq package [14].
4. The identification and classification of promoter-associated RNAs requires further bioinformatics analysis to associate their expression with known transcriptional start sites.

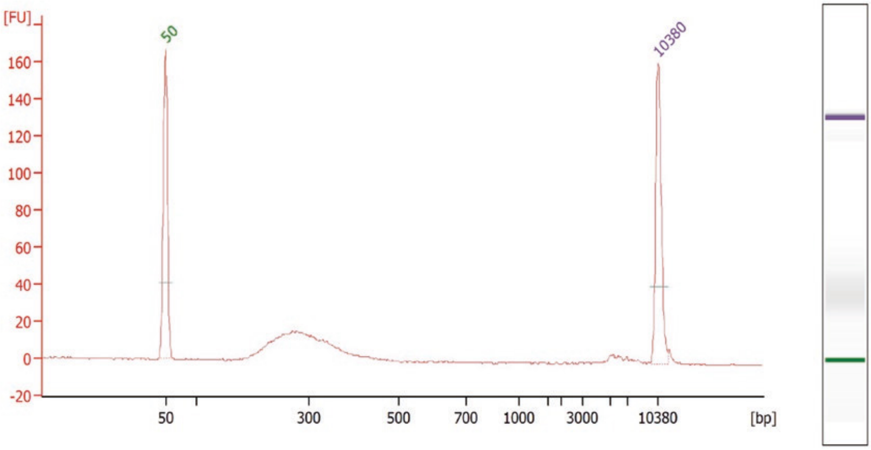


Fig. 2 Bioanalyzer result of a good-quality GRO-seq library is shown as an example. The fragments should be dispersed between 200 and 400 bp

4 Notes

1. It is important to equally expose the surface to swelling buffer. Rock the plates occasionally to prevent drying.
2. Resuspend nuclei by flicking the tube before adding the Lysis Buffer, and then invert tube several times.
3. In order to obtain better yields, it is advisable to use two separate reactions for NRO-rxn (two times of 5×10^6 nuclei).
4. Sarkosyl level and final concentration of CTP need to be adjusted according to the experimental system.
5. Having sarkosyl in NRO-mix causes the mixture to become very viscous. In order to obtain a homogenous mixture, cut the edge off a normal pipette tip and at least mix 15–20 times before placing the mixture at +30 °C.
6. If needed, combine two of the same reactions.
7. It is not recommended to use any salt containing buffers at this step. If you need to use salt containing buffer to obtain clear RNA pellet, you should wash the pellet at least two times with 70 % EtOH to remove the residual salt.
8. Upon fragmentation, it is expected to obtain a distribution of RNA fragments between 100 and 150 bp. The duration of the fragmentation reaction requires optimization and a bioanalyzer profile of different time points can provide the requisite information.

9. Handle the beads carefully because they do not form a rigid pellet. Allow 30–50 μl of SN to remain in the tube, in order to avoid losing your sample.
10. EDTA is needed to chelate divalent ions that catalyze cleavage of the RNA with heat.
11. You can check if the binding is at or near completion by spinning the beads down and removing the SN. Check the beads and supernatant with the Geiger counter. If the SN has 5–10× fewer counts than the beads, then the binding is likely complete.

Acknowledgment

We would like to specially thank Leighton Core for kindly sharing the GRO-seq protocol with us. RL is supported by the Fundação para a Ciência e Tecnologia de Portugal (SFRH/BD/74476/2010; POPH/FSE).

Reference

1. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63. doi:10.1038/nrg2484
2. TR C, JA S (2014) The noncoding RNA revolution-trashing old rules to forge new ones. Cell 157(1):77–94. doi:10.1016/j.cell.2014.03.008
3. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322(5909):1845–1848. doi:10.1126/science.1162228
4. Churchman LS, Weissman JS (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature 469(7330):368–373. doi:10.1038/nature09652
5. Kwak H, Fuda NJ, Core LJ, Lis JT (2013) Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science 339(6122):950–953. doi:10.1126/science.1229386
6. Leveille N, Melo CA, Rooijers K et al (2015) Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA. Nat Commun 6:6520. doi:10.1038/ncomms7520
7. Sun M, Gadad SS, Kim DS, Kraus WL (2015) Discovery, annotation, and functional analysis of long noncoding RNAs controlling cell-cycle gene expression and proliferation in breast cancer cells. Mol Cell 59(4):698–711. doi:10.1016/j.molcel.2015.06.023
8. Korkmaz, G. et al. Functional genetic screens for enhancer elements in the human genome using CRISPR–Cas9. Nat. Biotechnol. 34, 192–198 (2016). doi:10.1038/nbt.3450
9. Hah N, Murakami S, Nagari A et al (2013) Enhancer transcripts mark active estrogen receptor binding sites. Genome Res 23(8):1210–1223. doi:10.1101/gr.152306.112
10. Li W, Notani D, Ma Q et al (2013) Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature 498(7455):516–520. doi:10.1038/nature12210
11. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324(5924):218–223. doi:10.1126/science.1168978
12. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359. doi:10.1038/nmeth.1923
13. Heinz S, Benner C, Spann N et al (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38(4):576–589. doi:10.1016/j.molcel.2010.05.004
14. Anders S, Pyl PT, Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31(2):166–169. doi:10.1093/bioinformatics/btu638

Chapter 4

Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9

Adapted from Nature Biotechnology. 2016 Feb;34(2):192-8.

Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9

Gozde Korkmaz*, Rui Lopes*, Alejandro P Ugalde, Ekaterina Nevedomskaya, Ruiqi Han, Ksenia Myacheva, Wilbert Zwart, Ran Elkon & Reuven Agami

Systematic identification of noncoding regulatory elements has, to date, mainly relied on large-scale reporter assays that do not reproduce endogenous conditions. We present two distinct CRISPR-Cas9 genetic screens to identify and characterize functional enhancers in their native context. Our strategy is to target Cas9 to transcription factor binding sites in enhancer regions. We identified several functional enhancer elements and characterized the role of two of them in mediating p53 (*TP53*) and ERα (*ESR1*) gene regulation. Moreover, we show that a genomic CRISPR-Cas9 tiling screen can precisely map functional domains within enhancer elements. Our approach expands the utility of CRISPR-Cas9 to elucidate the functions of the noncoding genome

Enhancers are genomic elements that regulate transcription of distantly located genes through chromatin looping. They function as binding platforms for transcription factors and are characterized by specific chromatin modifications[1]. Recent studies have shown that genetic alterations can affect enhancer activity and contribute to tumorigenesis[2,3]. Moreover, transcription factors and other enhancer-associated proteins are frequently mutated in human tumors, and targeting these proteins with small-molecule inhibitors holds much therapeutic potential[1,4]. It is estimated that the human genome contains >500,000 putative enhancers, a staggering number that poses a major challenge for the identification of functional regulatory elements. Current methods to systematically identify enhancers are based on massively parallel reporter sequencing. However, the intrinsically artificial nature of these methods is likely to have some effect on their ability to delineate and assess the activity of endogenous enhancers. The recent development of CRISPR-Cas9 (clustered, regularly interspaced, short palindromic repeats (CRISPR) and the CRISPR-associated protein 9 (Cas9)) technology has opened unprecedented opportunities for genome-wide targeted editing in human cells[5]. Previous functional genetic screens using CRISPR-Cas9 have mainly been restricted to protein-coding genes[6,7]. Here, we apply this technology to identify endogenous enhancer elements and to characterize the domains that are essential for their activity.

As a proof-of-principle demonstration, we focused on two transcription factors, p53 and ERα, which play key roles in cancer initiation and progression. p53 is known as the 'guardian of the genome', and is mutated in more than 50% of all human tumors[8]. Upon oncogene activation, one major function of p53 is to activate an irreversible cell-cycle arrest program named oncogene-induced senescence (OIS). OIS is a powerful tumor-suppressive mechanism; somatic mutations in p53, or in other components of its pathway, can overcome OIS and lead to tumorigenesis[9]. ERα is an estrogen-activated transcription factor that has a mitogenic role in breast cancer cells. The standard of care for ERα-positive breast tumors is treatment with selective ERα modulators and aromatase inhibitors. However, many tumors relapse after treatment and most of them still express *ERα* (also known as *ESR1*)[10]. Recently, both p53 and ERα have been shown to directly bind genomic regions that are characterized largely by features of distal-enhancer regions[11,12]. This evidence suggests that the identification of p53- and ERα-bound enhancers and their target genes could be instrumental for diagnostics and therapeutics of cancer.

Initially, we set out to establish a genetic screen for p53-bound enhancers that are required for OIS (**Fig. 1a**). To build a CRISPR-Cas9 single guide RNA (sgRNA) library, we followed a strategy that enabled us to target ≈90% of p53-bound enhancers (**Fig. 1b** and **Supplementary Table 1**). We cloned the sgRNAs into lentiCRISPRv2 vector[13] using a pooled strategy and generated a lentiviral library (CRISPR-p53-enhLib). We performed our screen in human BJ cells containing tamoxifen-inducible HRAS^{G12V} (BJ-RAS^{G12V}), which are a well-characterized cell model of OIS[14,15]. Accordingly, we transduced cells with three independent lentiviral pools of CRISPR-p53-enhLib, as well as with a nontargeting sgRNA pool (negative control) (**Fig. 1c**). After 4 weeks of culturing, we harvested the cells and performed next-generation sequencing to identify the sgRNAs present in the populations (**Fig. 1c** and **Supplementary Fig. 1a**).

Our screen detected eight substantially enriched sgRNAs ($q < 0.1$) in the RAS^{G12V}-induced cell populations (**Fig. 1d** and **Supplementary Table 2**). Notably, two independent sgRNAs targeted a putative enhancer located ~10 kb upstream of *CDKN1A* (formerly known as *p21*), which is a key effector of p53-dependent OIS (**Fig. 1e**; p53^{enh3507}). Another top-scoring sgRNA mapped to a known p53-responsive

*These authors contributed equally to this work. Correspondence should be addressed to R.E. (r.elkon@nki.nl) and R.A. (r.agami@nki.nl).

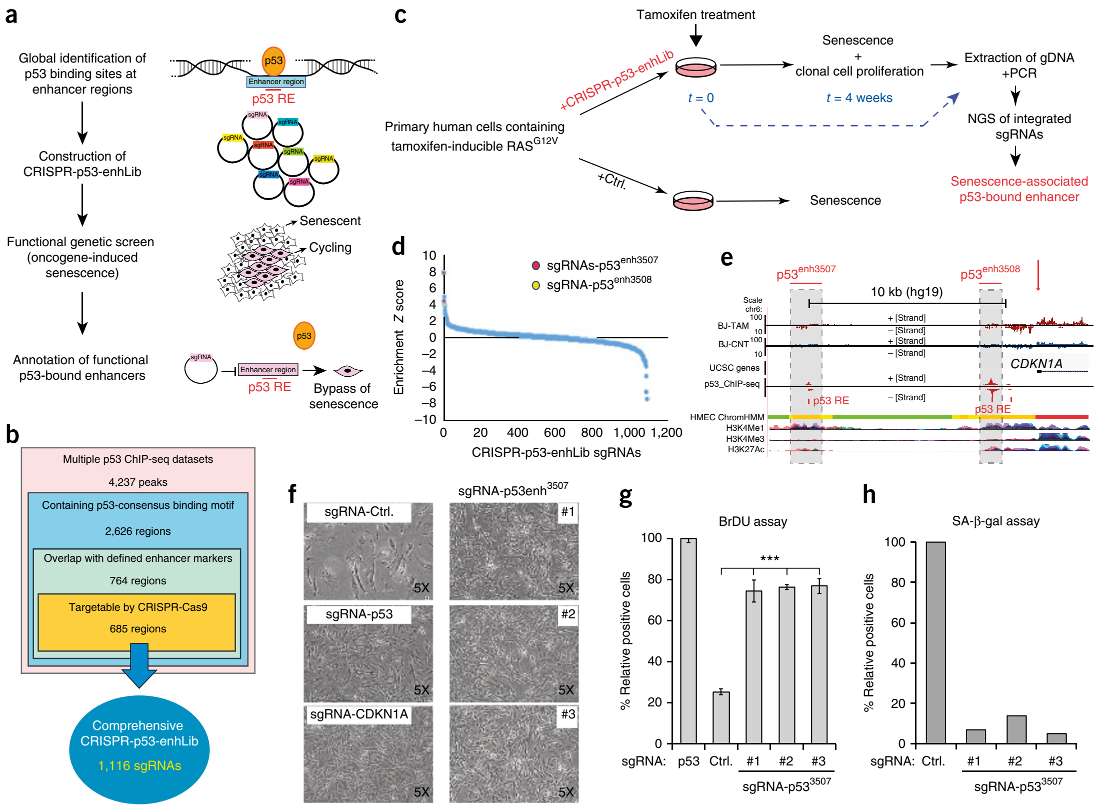


Figure 1 A comprehensive CRISPR-Cas9 genetic screen identifies p53-bound enhancers required for OIS. (**a**) Screening strategy for detecting p53-bound enhancers required to elicit OIS. RE, responsive element (**b**) Summary of pipeline used to establish a genome-wide CRISPR-Cas9–based library targeting p53-bound (termed CRISPR-p53-enhLib). (**c**) The functional genetic screen procedure. NGS, next-generation sequencing. (**d**) Enrichment score calculated for each sgRNA vector based on its prevalence in the pool, harvested after 4 weeks of tamoxifen (TAM) treatment (HRAS^{G12V} induction), relative to its prevalence in the untreated pool. The plot shows the distribution of standardized enrichment scores (*Z*-scores) for the entire CRISPR-p53-enhLib library. A red dot indicates two independent sgRNAs targeting p53^{enh3507}; the yellow dot, an sgRNA that targets p53^{enh3508}. n.s., not significant. (**e**) Genomic tracks for p53 binding events, histone modifications (based on publicly available data sets) and for transcriptional activity measured by GRO-seq in induced BJ-RAS^{G12V} (BJ-TAM) and noninduced BJ-RAS^{G12V} cells (BJ-CNT). The colors in the chromatin hidden Markov model track represent predicted chromatin function as follows: orange, strong enhancer; yellow, weak enhancer; red, promoter; green, transcriptional activity. (**f**) Light microscopy images of cell populations transduced with the indicated sgRNA vectors. Images were taken after 15 d of HRAS^{G12V} induction. (**g**) Proliferation of the various CRISPR-Cas9–transduced BJ-RAS^{G12V} cells was quantified using a BrdU assay. $N = 2$; for each condition, at least 150 cells were counted. $***P < 0.005$, two-tailed Student's *t*-test. For every condition, percentage of BrdU-positive cells was normalized to p53KO cells. (**h**) Senescence induction was quantified using senescence-associated (SA) β-gal assay. For every condition, percentage of β-gal-positive cells was normalized to Ctrl cells. Error bars, mean ± s.d.

element that is located proximal to the transcription start site of *CDKN1A* (**Fig. 1e**; p53^{enh3508}). To validate the results of the screen, we repeated the OIS experiment using individual sgRNAs. In this assay, we included sgRNAs targeting the coding region of *p53* and *CDKN1A* as positive controls, and a nontargeting sgRNA as a negative control. As additional negative controls, we tested three sgRNAs that did not show significant enrichment in our screen (p53^{enh1646}, p53^{enh2736} and p53^{enh3962}). These experiments validated four out of eight original hits, and identified three different enhancers that are required for OIS: p53^{enh3507}, p53^{enh3508} and p53^{enh1396} (**Fig. 1f** and **Supplementary Fig. 1b**,**c**). Whereas p53^{enh3507} and p53^{enh3508} presumably co-regulate *CDKN1A* expression, p53^{enh1396} is located in an intron of the long noncoding RNA RP11-382A20. Of note, none of the negative-control sgRNAs caused OIS bypass, indicating the robustness of our assay. (**Supplementary Fig. 1b**,**c**).

We focused subsequent experiments on p53^{enh3507}, which is a putative enhancer region that has not previously been functionally characterized in human cells, and on p53^{enh3508}, whose role in mediating p53-dependent *CDKN1A* regulation has been previously documented[16]. We performed BrdU labeling and senescence-associated β-galactosidase (β-gal) assays and found that sgRNA-p53^{enh3507} and sgRNA-p53^{enh3508} caused OIS bypass and continuous cell proliferation (**Fig. 1g**,**h** and **Supplementary Fig. 2a**). Next, we assessed the enhancer capacity of the p53^{enh3507} region by cloning it into a reporter vector and verified that it strongly induces transcription (**Fig. 2a**). We also observed a substantial reduction in enhancing activity upon

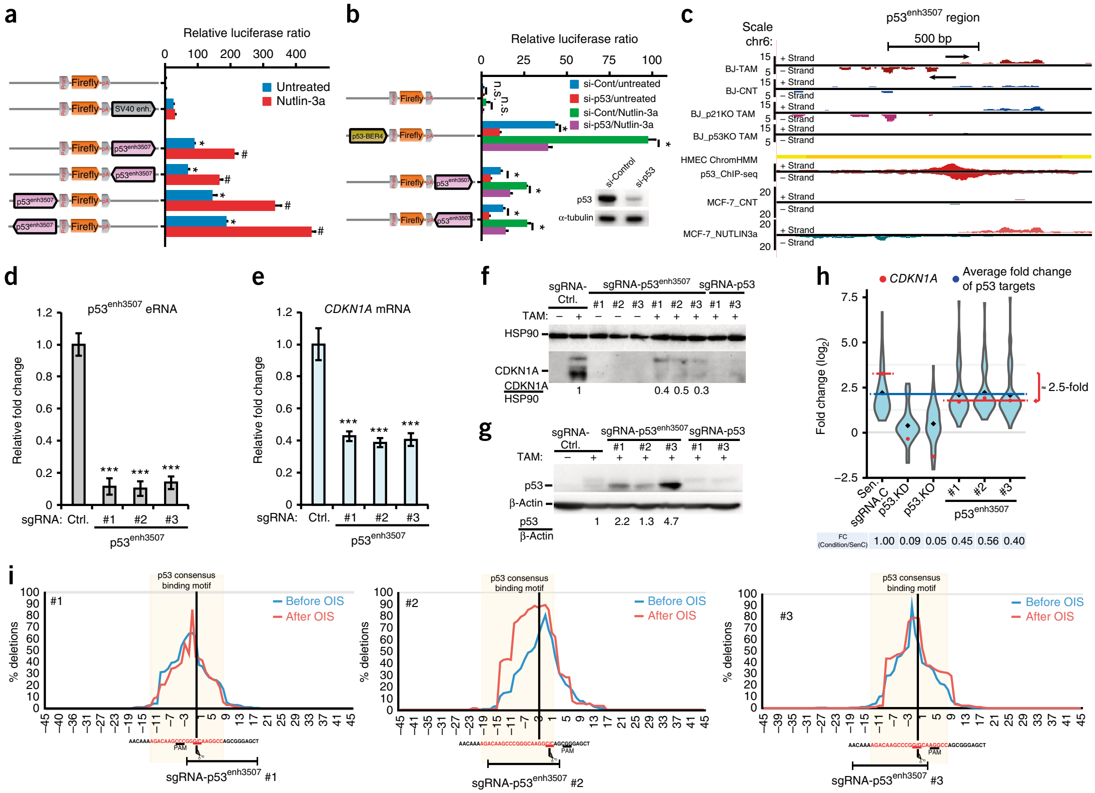


Figure 2 p53^{enh3507} is a p53-dependent enhancer region that regulates *CDKN1A* expression. (**a**) MCF-7 cells were transfected with the indicated reporter vectors, treated with Nutlin-3a 5–10 h later, and harvested 25–30 h after treatment. The relative luciferase activities (firefly/*Renilla*) were normalized to the control (Ctrl.) reaction. *P*-values for luciferase assay were calculated by two-tailed Student's *t*-test. $*P < 0.005$, relative to empty vector; $\#P < 0.005$, relative to untreated matching sample. (**b**) The same assay as in **a**, only that cells were co-transfected with control, or p53-targeting short interfering RNA (si-Cont.; si-p53). A reporter vector containing the enhancer region p53-BER4 was used as a positive control for p53-dependency. The efficiency of p53 knockdown was determined by immunoblot analysis. *P*-values for luciferase assay were calculated by two-tailed Student's *t*-test. $*P < 0.01$, relative to empty vector. (**c**) GRO-seq analysis detected strong induction of eRNA expression at p53^{enh3507} upon RAS^{G12V} induction in BJ cells. This induction was completely abolished by p53-KO but not affected by CDKN1A-KO. The knockouts of p53 and CDKN1A were verified by western blot analysis (**Fig. 2f** and **Supplementary Fig. 3f**). Activation by Nutlin-3a treatment in MCF-7 cells resulted in a strong induction of eRNA in this region. (**d**,**e**) qRT-qPCR measurements of either eRNAs transcribed from the p53^{enh3507} region (**d**) or mRNAs of *CDKN1A* (**e**). $n = 3$; $***P < 0.005$, $*P < 0.05$, two-tailed Student's *t*-test. (**f**,**g**) Immunoblot analysis for CDKN1A (**f**) and p53 (**g**) proteins in BJ-RAS^{G12V} cells transduced as indicated, and treated with tamoxifen (TAM) for 12–15 d. HSP90 and β-Actin protein levels are shown as a loading control. (**h**) Using RNA-seq, we identified a set of 54 known direct target genes of p53 that were induced by at least twofold upon HRAS^{G12V} activation in BJ cells (Sen.C), and examined the effect of various vector transductions on the induction of this set of genes: Sen.sgRNAs.c, control sgRNA; p53.KD, knockdown of p53 using siRNA; p53. KO, knockout of p53 using CRISPR-Cas9 which targets p53; #1, #2 and #3: three independent sgRNA vectors targeting the p53 binding site within p53^{enh3507}. Violin plots show the distribution of fold-induction of the set of 54 direct targets of p53 in each condition. Blue diamond indicates average induction of p53 targets; red dot indicates the level of *CDKN1A* mRNA induction. (**i**) We deep-sequenced a genomic region of 100 nt centered at the p53-consensus binding site of p53^{enh3507} from control and HRAS^{G12V}-induced BJ cells transduced with the indicated sgRNAs. We calculated the prevalence of deletions that occurred at each position within this interval (relative to the total number of reads that contained any deletion). The p53 consensus motif and the location of the sgRNA-mediated CRISPR-Cas9 endonuclease cut are indicated. Error bars, mean ± s.d.

siRNA-mediated knockdown of p53 (**Fig. 2b**). We obtained similar results for the p53^{enh3508} enhancer element (**Supplementary Fig. 2b**,**c**). Taken together, our results demonstrate that p53^{enh3507} controls OIS, and its transcription-enhancing activity is p53-dependent.

Active enhancer regions produce enhancer-associated RNAs (eRNAs), whose expression levels correlate with enhancer activity[17]. Indeed, we performed global run-on sequencing (GRO-seq) in BJ-RAS^{G12V} cells and detected strong induction of eRNA expression from the p53^{enh3507} region upon activation of OIS (**Fig. 2c**). CRISPR-Cas9–mediated knockout of *p53*, but not of *CDKN1A*, completely abolished eRNA expression from this region, indicating p53-dependent regulation of p53^{enh3507} (**Fig. 2c**). GRO-seq data of MCF-7 cells treated with Nutlin-3a suggest that this region is an active p53-responsive enhancer in different types of human cells[18] (**Fig. 2c**). Next, we measured eRNA expression from the p53^{enh3507} region and found that all three sgRNA-p53^{enh3507} caused about a tenfold reduction in eRNA

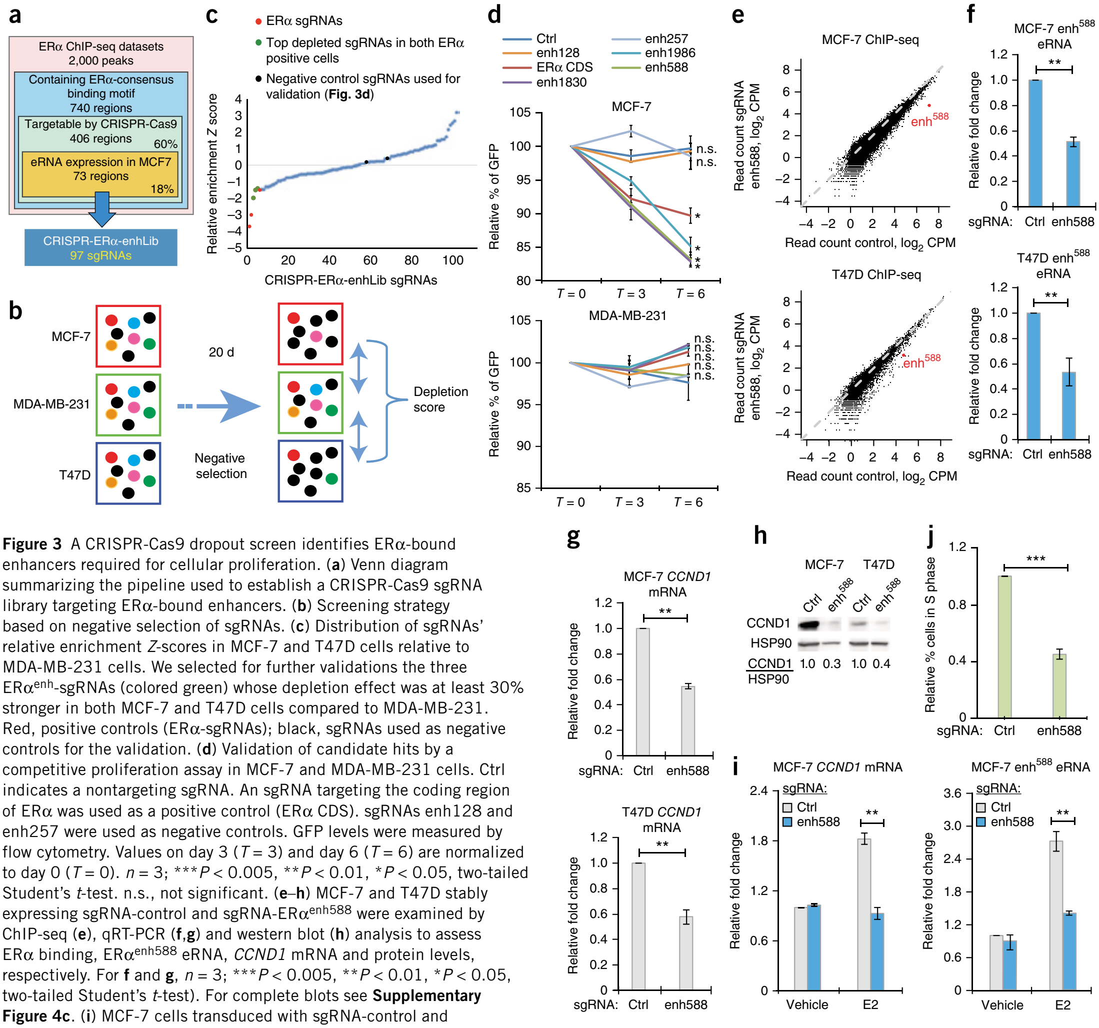


Figure 3 A CRISPR-Cas9 dropout screen identifies ERα-bound enhancers required for cellular proliferation. (**a**) Venn diagram summarizing the pipeline used to establish a CRISPR-Cas9 sgRNA library targeting ERα-bound enhancers. (**b**) Screening strategy based on negative selection of sgRNAs. (**c**) Distribution of sgRNAs' relative enrichment *Z*-scores in MCF-7 and T47D cells relative to MDA-MB-231 cells. We selected for further validations the three $ER\alpha^{enh}$-sgRNAs (colored green) whose depletion effect was at least 30% stronger in both MCF-7 and T47D cells compared to MDA-MB-231. Red, positive controls (ERα-sgRNAs); black, sgRNAs used as negative controls for the validation. (**d**) Validation of candidate hits by a competitive proliferation assay in MCF-7 and MDA-MB-231 cells. Ctrl indicates a nontargeting sgRNA. An sgRNA targeting the coding region of ERα was used as a positive control (ERα CDS). sgRNAs enh128 and enh257 were used as negative controls. GFP levels were measured by flow cytometry. Values on day 3 ($T = 3$) and day 6 ($T = 6$) are normalized to day 0 ($T = 0$). $n = 3$; $***P < 0.005$, $**P < 0.01$, $*P < 0.05$, two-tailed Student's *t*-test. n.s., not significant. (**e**–**h**) MCF-7 and T47D stably expressing sgRNA-control and sgRNA-$ER\alpha^{enh588}$ were examined by ChIP-seq (**e**), qRT-PCR (**f**,**g**) and western blot (**h**) analysis to assess ERα binding, $ER\alpha^{enh588}$ eRNA, *CCND1* mRNA and protein levels, respectively. For **f** and **g**, $n = 3$; $***P < 0.005$, $**P < 0.01$, $*P < 0.05$, two-tailed Student's *t*-test). For complete blots see **Supplementary Figure 4c**. (**i**) MCF-7 cells transduced with sgRNA-control and sgRNA-$ER\alpha^{enh588}$ were depleted of estrogens for 72 h and then stimulated with 17β-estradiol for 24 h. Relative expression levels of $ER\alpha^{enh588}$ eRNA and CCND1 mRNA were measured by qRT-PCR. $n = 3$; $***P < 0.005$, $**P < 0.01$, $*P < 0.05$, two-tailed Student's *t*-test. (**j**) S-phase entry of sgRNA-control and $ER\alpha^{enh588}$-transduced MCF-7 cells was assayed by BrdU incorporation and measured by flow cytometry. $n = 4$; for each condition 10,000 events were recorded. $***P < 0.005$, $**P < 0.01$, $*P < 0.05$, two-tailed Student's *t*-test. Error bars, mean ± s.d.

expression (**Fig. 2d**), and a corresponding ~2.5-fold reduction in *CDKN1A* mRNA (**Fig. 2e**) and protein (**Fig. 2f**) levels. Importantly, the reduction of $p53^{enh3507}$ eRNA and *CDKN1A* mRNA expression occurred in conditions of elevated p53 protein levels and continuous $HRAS^{G12V}$ expression (**Fig. 2g** and **Supplementary Fig. 3a**). Finally, we performed RNA-seq in $BJ\text{-}RAS^{G12V}$ cells and confirmed that the effect of sgRNA-$p53^{enh3507}$ is specific to *CDKN1A* (**Fig. 2h**; red dot). In contrast, downregulation of p53 significantly ($P = 6.1 \times 10^{-13}$ for p53KD and $P = 8.3 \times 10^{-8}$ for p53KO; Wilcoxon's test) reduced the expression of the majority of its target genes (**Fig. 2h**; blue dot). These results indicate that sgRNA-$p53^{enh3507}$ disrupts the enhancing activity of this region and decreases the activation of *CDKN1A* upon OIS.

Cas9-nuclease activity generates DNA double-strand breaks that result in deletions and insertions in the vicinity of the sgRNA recognition site[19]. Therefore, we examined the spectrum of deletions caused by sgRNA-$p53^{enh3507}$ in control and induced $BJ\text{-}RAS^{G12V}$ cells. We found that sgRNA-$p53^{enh3507}$ caused deletions ranging from 1–15 nucleotides (**Fig. 2i**). The mutations were restricted to the p53 binding site following OIS induction, indicating that small deletions in key noncoding regulatory sequences are sufficient to cause a phenotypic change *in vivo* (**Fig. 2i**). We also tested the effect of CRISPR-Cas9–induced mutations on the activity of $p53^{enh3507}$ region in a reporter vector. In all cases, p53-dependent enhancer activity was abolished by the mutations (**Supplementary Fig. 3c**), indicating that an intact p53 binding site is indispensable for the $p53^{enh3507}$ enhancing function.

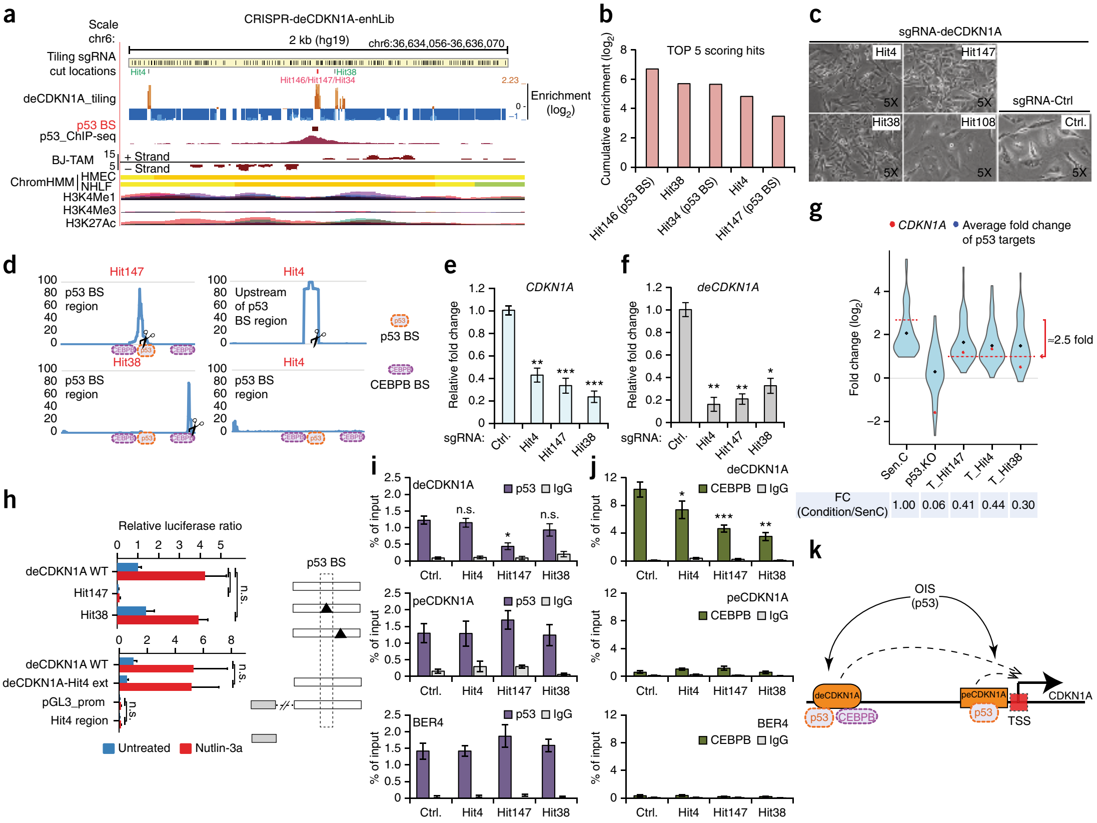


Figure 4 CRISPR-Cas9 tiling screen uncovers novel elements required for enhancer function. (**a**) Top panel, a snapshot of the genomic region surrounding *deCDKN1A*. Each black line indicates one sgRNA (total 197 sgRNAs covering ~2 kb). At the middle, the result of the CRISPR-deCDKN1A-enhLib screen is shown. Enrichment score was calculated for each sgRNA based on the ratio between its (normalized) prevalence in the HRAS^{G12V}-induced and control BJ cells. Blue, OIS-depleted sgRNAs; orange, OIS-enriched ones; green, sgRNAs targeting novel candidate regulatory elements; red, positive control sgRNAs that target the p53 binding site (BS). UCSC tracks of p53-ChIP and chromatin annotation are as in **Figure 1e**. HMEC, human mammary epithelial cell; NHLF, human lung fibroblast; ChromHMM, chromatin state segmentation by HMM; H3K4Me1, histone H3 lysine 4 monomethylation; H3K4Me3, histone H3 lysine 4 trimethylation; H3K27Ac, histone H3 lysine 27 acetylation. (**b**) Top five hits from the tiling screen and their cumulative enrichment from three independent screens. (**c**) Light microscopy images of cells transduced with the indicated sgRNAs. (**d**) Sequencing spectrums of cleavage sites of top tiling screen hits. The spectrum of Hit38 was very narrow and resembled the one induced by Hit147 that targets the p53 BS. The spectrum of Hit4 was dominated by a deletion of ~20 nt. As control, we also sequenced the p53 BS region in cells targeted with Hit4. As the cleavage induced by Hit4 is located outside this genomic interval, there was no marked peak in this region. (**e**,**f**) qRT-PCR of CDKN1A mRNA (**e**) and deCDKN1A-eRNA (**f**) of BJ-RAS^{G12V} cells transduced with the indicated sgRNAs, and treated as in **Figure 2d**. $n = 3$; $***P < 0.005$, $**P < 0.01$, $*P < 0.05$, two-tailed Student's *t*-test. (**g**) RNA-seq result depicted as violin plots to show the distribution of p53 targets genes in each condition. Blue diamond indicates average induction of p53 targets; red dot indicates the level of *CDKN1A* mRNA induction. (**h**) Luciferase assay of MCF-7 cells transfected with the indicated reporter vectors and induced with Nutlin-3a. The relative luciferase activities (firefly/*Renilla*) were normalized to the control reaction. All *P*-values for luciferase assay were calculated by Student's *t*-test. (**i**,**j**) Bar graphs of p53 (**i**) and CEBPB (**j**) binding at deCDKN1A, peCDKN1A and p53-BER4 regions by ChIP-qPCR. BJ-RAS^{G12V} cells were infected with either sgRNA targeting Ctrl or each newly identified sgRNAs and treated with tamoxifen (TAM) for 12–15 d. $n = 3$; n.s., not significant, $***P < 0.005$, $**P < 0.01$, $*P < 0.05$, two-tailed Student's *t*-test. (**k**) A schematic model showing that p53 binding at both distal and proximal (deCDKN1A and peCDKN1A, respectively) regions is required for sustained high level of CDKN1A, and for OIS. Error bars, mean ± s.d.

Interestingly, concurrent disruption of p53^{enh3507} and p53^{enh3508} resulted in further reduction of *CDKN1A* expression, suggesting an independent regulation by both enhancer elements (**Supplementary Fig. 3b**; left panel). In comparison, downregulation of p53 resulted in even lower levels of *CDKN1A*, possibly due to indirect effects on *CDKN1A* expression[18]. However, BrdU and β-gal assays showed no additional effect by the combined sgRNAs compared with the singles, indicating that inactivation of one enhancer is sufficient to complete a phenotypic alteration (**Supplementary Fig. 3d**,**e**). Collectively, these results demonstrate that p53^{enh3507} is an endogenous enhancer of CDKN1A and disruption of this region causes bypass of senescence in p53-WT cells. Moreover, cooperative action of p53^{enh3507} (distal enhancer of *CDKN1A—deCDKN1A*) and p53^{enh3508} (proximal enhancer of *CDKN1A—peCDKN1A*) is required to activate *CDKN1A* expression and initiate OIS.

To demonstrate the generalizability of our screening approach, we designed a dropout screen to identify novel ERα-bound enhancers. First, we selected two breast cancer cell lines (MCF-7 and T47D) that require ERα for cell proliferation, and one (MDA-MB-231) that lacks ERα expression. We generated a sgRNA library to target 73 ERα binding sites according to the strategy shown in **Figure 3a** (CRISPR-ERα-enhLib; **Supplementary Table 3**). For this library, we used eRNA expression measured by GRO-seq in MCF-7 cells as a criteria for active enhancers[12]. As positive controls, we included in the library three sgRNAs targeting the coding region of *ERα*. Accordingly, we transduced the three different cell lines with the CRISPR-ERα-enhLib and allowed the cells to proliferate for 20 d (**Fig. 3b**). After identifying the sgRNAs present in the cell populations, we ranked the hits by negative effect on cell proliferation and selected sgRNAs that were strongly depleted (at least 30%) in both MCF-7 and T47D compared to MDA-MB-231 cells. Using these criteria, we identified two positive controls that target the *ERα* and three candidate sgRNAs (ERα^{enh588}, ERα^{enh1830} and ERα^{enh1986}) that affect the proliferation of MCF-7 and T47D cells (**Fig. 3c** and **Supplementary Table 4**). We validated the candidates with a competitive proliferation assay in MCF-7 and MDA-MB-231 cells (**Fig. 3d**). As expected, sgRNA-ERα reduced the proliferation of MCF-7 cells but had no effect on MDA-MB-231 cells. Remarkably, all three candidates identified in the screen also significantly decreased the proliferation of only MCF-7 cells. In contrast, two sgRNAs (ERα^{enh128} and ERα^{enh257}) that were not depleted in the screen, and a nontargeting sgRNA control, had no significant effect on cell proliferation. These results indicate that our CRISPR-Cas9 dropout screening approach is specific, and led to the identification of three novel enhancers that are required for breast cancer cell proliferation.

One of the validated candidates, ERα^{enh588}, has not been endogenously characterized to date. A ChIA-PET (chromatin-interaction analysis by paired-end tag) study in MCF-7 cells has previously identified ERα^{enh588} as a hotspot for ERα binding[20]. Analyses of this data set showed that ERα^{enh588} interacts with the promoter region of Cyclin D1 (*CCND1*) (**Supplementary Fig. 4a**), suggesting that ERα^{enh588} is a putative regulator of *CCND1* expression by ERα. *CCND1* oncogene plays a central role in cell-cycle progression and is overexpressed in more than 50% of breast tumors[21]. We started by cloning ERα^{enh588} WT region in a reporter vector and verified that it has strong ERα–dependent transcription-enhancing activity, since mutations in the ERα binding site completely abolish the enhancer activity and response to 17β-estradiol (**Supplementary Fig. 4b**). Next, we examined endogenous ERα binding at ERα^{enh588} region by ChIP-seq and observed a substantial decrease in both MCF-7 and T47D cells expressing sgRNA-ERα^{enh588} (**Fig. 3e**). Because *CCND1* is a putative target of ERα^{enh588}, we verified the endogenous expression of ERα^{enh588} eRNAs and *CCND1* mRNA and protein in MCF-7 and T47D cells transduced with sgRNA-ERα^{enh588}. Reassuringly, we found that the expression of eRNA, mRNA and protein is significantly decreased ($P < 0.01$; about twofold) in both cell lines (**Fig. 3f–h**). These results indicate that ERα^{enh588} is a bona fide enhancer element that regulates the expression of *CCND1*. Next, we assessed the dependency of ERα^{enh588} and *CCND1* endogenous expression on estrogen signaling by qPCR (**Fig. 3i**). As expected, we confirmed that both ERα^{enh588} eRNA and *CCND1* mRNA are upregulated in MCF-7 cells upon treatment with 17β-estradiol. However, we found that sgRNA^{enh588} severely compromises the induction of eRNA expression and completely abolishes *CCND1* mRNA activation in MCF-7 cells. These results suggest that the activation of *CCND1* expression by estrogen in breast cancer cells requires a fully active ERα^{enh588} enhancer element. Finally, as CCND1 is a crucial component of the G1-S phase transition, we examined the phenotypic outcome of disrupting ERα^{enh588} activity. **Figure 3j** shows that MCF-7 cells transduced with sgRNA^{enh588} display a ~2.5-fold reduction in S-phase entry, compared to control-transduced cells, due to decreased *CCND1* expression.

The genetic code that enables enhancer activity is poorly understood. We reasoned that each enhancer is likely to contain multiple regulatory elements, and adopted CRISPR-Cas9 technology to pinpoint critical domains of enhancers. For that purpose, we performed a pooled high-throughput genetic tiling screen to identify additional elements, apart from the p53 binding site, that are required for the p53^{enh3507} region to regulate *CDKN1A* expression and function in OIS.

We targeted a genomic region of ~2 kb centered at the p53 binding site of *deCDKN1A* (**Fig. 4a**). We identified protospacer-adjacent motifs (PAMs) within this region and designed a library of 197 sgRNAs (CRISPR-deCDKN1A-Lib) that direct CRISPR-mediated cleavage every 10 bp on average (**Supplementary Table 5**). We performed an OIS screen with the CRISPR-deCDKN1A-Lib following the same procedure described in **Figure 1c**, which identified five enriched sgRNAs (**Fig. 4b** and **Supplementary Table 6**). Three of the most-enriched hits targeted the p53 binding site (Hit146, Hit147 and Hit34), and thus served as positive controls and as indicators of the robustness of the assay. The other two hits, Hit4 and Hit38, targeted novel candidate regulatory domains located 0.9 kb upstream and 0.1 kb downstream of the p53 binding site, respectively (**Fig. 4a**, middle panel). For validation, we used individual sgRNAs and confirmed that Hit4 and Hit38, similarly to the positive control Hit147, caused bypass of OIS (**Fig. 4c**). We analyzed the spectrum of mutations generated by these two sgRNAs and verified that it was very similar to the one caused by sgRNAs that directly targeted the p53 binding site (**Figs. 2i** and **4d**). Importantly, sgRNA-Hit4 and sgRNA-Hit38 caused genomic alterations that did not overlap with the p53-binding site within this enhancer region, indicating that they disrupt other enhancer domains required for OIS (**Fig. 4d**).

To characterize the function of these two regulatory domains, we first examined the expression of *CDKN1A* mRNA and *deCDKN1A* eRNA in BJ-RAS^{G12V} cells transduced with sgRNA-Hit4 and sgRNA-Hit38. We observed a significant ($P < 0.005$, and < 0.01; and $P < 0.01$ and < 0.05, respectively) reduction in both mRNA (**Fig. 4e**) and eRNA (**Fig. 4f**) expression levels, which was similar in magnitude to that of the positive control sgRNA-Hit147. Gene expression profiling by RNA-seq also supported a specific effect of these two sgRNAs on *CDKN1A* expression (**Fig. 4g**). However, unlike mutations in the p53 binding site, deletion of the Hit4 or Hit38 region did not affect enhancer function in a reporter assay (**Fig. 4h**), suggesting that these domains only have a functional role in their endogenous context. Next, we analyzed endogenous p53 binding to the *deCDKN1A* region using ChIP-qPCR. As expected, Hit147 significantly ($P < 0.05$) reduced p53 binding to this region, whereas both Hit4 and Hit38 had no significant effect (**Fig. 4i**). For this experiment, we used *peCDKN1A* and p53BER4 regions as negative controls (**Fig. 4i**). We used PROMO[22] to identify transcription factors that potentially bind to these regulatory domains and identified a perfect matching consensus binding site for CEBPB at the cleavage site of Hit38 (**Supplementary Fig. 5**). In contrast, no putative transcription factor binding site was predicted for the Hit4 region. Accordingly, we analyzed endogenous CEBPB binding to *deCDKN1A* by ChIP-qPCR and observed a marked reduction in BJ-RAS^{G12V} cells transduced with sgRNA-Hit38 compared to control cells (**Fig. 4j**). These results suggest that the DNA element targeted by sgRNA-Hit38 contributes to CEBPB recruitment to *deCDKN1A* and to its function in OIS (**Fig. 4k**). In addition, Hit147 also decreased CEBPB binding, possibly due to an additional CEBPB binding site located adjacent to the p53 binding site. Also, Hit4 had a weak but significant ($P < 0.005$, < 0.01 and < 0.05) effect on CEBPB binding at the *deCDKN1A* region (**Fig. 4j**). Finally, we tested whether Hit4 and Hit38 had a cooperative effect on *CDKN1A* expression but, unlike *deCDKN1A* and *peCDKN1A*, this proved not to be the case (**Supplementary Fig. 3b**, right panel). Altogether, our results indicate that a CRISPR-Cas9 tiling strategy can precisely pinpoint regulatory domains within enhancer regions.

In summary, we present CRISPR-Cas9–based screens to identify and characterize functional enhancers in human cells. In total, we identified six enhancer elements that potentially control cell proliferation, and characterized two of them in detail—one regulates *CDKN1A* activation during OIS and the other mediates *CCND1* expression in response to ERα signaling (**Figs. 2** and **3**). We observed different rates of validation of candidates between the two genetic screens presented here, with the ERα-bound enhancer showing higher specificity. We speculate that this difference might be related to the intrinsic nature of the two screens (enrichment vs. dropout) or to the selection procedure of candidates (selecting ERα-bound enhancers based on eRNA expression). Of note, none of the control candidates from either screen showed any phenotypic activity. This evidence suggests that our screening approach has comparable specificity and sensitivity to genetic screens of protein-coding genes performed to present[23].

Recently, a dCas9-LSD1 fusion has been proposed to annotate native enhancers[24], yet the sensitivity and specificity of this tool has never been tested in large-scale genetic screens. Our method expands the utility of the CRISPR-Cas9 tool beyond the coding genome and can be applied to systematically identify functional enhancers bound by different sequence-specific transcription factors. At the present date, all CRISPR-Cas9–based systems require a PAM motif to direct DNA cleavage, and therefore cannot guarantee full coverage of the entire human genome. In our approach, about 90% and 60% of the candidate p53- and ERα-bound enhancers were targeted, respectively, but this rate might be different for other transcription factors. However, the use of CRISPR-Cas9 nucleases with altered or reduced PAM specificities[25] can increase the coverage of our approach to target enhancers in large-scale genetic screens. The selection of candidate enhancer regions for this study was based on ENCODE, ChIP-seq and GRO-seq data sets, but in principle our approach does not require prior data on transcription factor binding site. As an alternative, active enhancer regions can be detected by transcriptomic profiling of eRNA expression and targeted in an unbiased fashion using a CRISPR-Cas9 tiling approach such as we present here and as others have recently demonstrated for the enhancer of *BCL11A*[26].

Our study shows that CRISPR-Cas9 technology is a robust tool to identify and characterize functional enhancers in an unbiased fashion. We envision that our approach will be widely used to unravel the function of the noncoding portion of the human genome under both normal and pathological conditions.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. GEO: GSE75627 (RNA-seq); GSE75779 (ChIP-seq).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to acknowledge K. Rooijers for help in bioinformatic analyses, insightful ideas and critical discussions. We thank W. de Laat, P. Krijger, B. Evers, M. Amendola and R. Maia for reagents and technical assistance. We also thank all members of the Agami group for helpful discussions. And we thank F. Zhang (MIT) for the gift of plasmid vector lentiCRISPRv2 and D. Root (MIT) for pLX304-GFP. R.L. is a fellow of the Fundação para a Ciência e Tecnologia de Portugal (SFRH/BD/74476/2010; POPH/FSE). This work was supported by funds of The Human Frontier Science Program (LT000640/2013) to A.P.U., and enhReg ERC-AdV and NWO (NGI 93512001/2012) to R.A.

AUTHOR CONTRIBUTIONS

G.K., R.L. and R.A. conceived the project. G.K. and R.L. performed most of the experiments. A.P.U. did all luciferase assays. W.Z. and E.N. carried out and analyzed the ChIP-Seq data, respectively. R.H. helped during the library preparation. K.M. did cloning and validation of candidates from the screens. R.E. conducted all the bioinformatics analyses. G.K., R.L. and R.A. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Sexton, T. & Cavalli, G. The role of chromosome domains in shaping the functional genome. *Cell* **160**, 1049–1059 (2015).
2. Mansour, M.R. *et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377 (2014).
3. Vahedi, G. *et al.* Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature* **520**, 558–562 (2015).
4. Shi, J. & Vakoc, C.R. The mechanisms behind the therapeutic activity of BET bromodomain inhibition. *Mol. Cell* **54**, 728–736 (2014).
5. Cho, S.W., Kim, S., Kim, J.M. & Kim, J.S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.* **31**, 230–232 (2013).
6. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
7. Wang, T., Wei, J.J., Sabatini, D.M. & Lander, E.S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).
8. Muller, P.A. & Vousden, K.H. p53 mutations in cancer. *Nat. Cell Biol.* **15**, 2–8 (2013).
9. Schmitt, C.A. *et al.* A senescence program controlled by p53 and p16INK4a contributes to the outcome of cancer therapy. *Cell* **109**, 335–346 (2002).
10. Beelen, K., Zwart, W. & Linn, S.C. Can predictive biomarkers in breast cancer guide adjuvant endocrine therapy? *Nat. Rev. Clin. Oncol.* **9**, 529–541 (2012).
11. Melo, C.A. *et al.* eRNAs are required for p53-dependent enhancer activity and gene transcription. *Mol. Cell* **49**, 524–535 (2013).
12. Li, W. *et al.* Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516–520 (2013).
13. Sanjana, N.E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
14. Drost, J. *et al.* BRD7 is a candidate tumour suppressor gene required for p53 function. *Nat. Cell Biol.* **12**, 380–389 (2010).
15. Voorhoeve, P.M. *et al.* A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors. *Cell* **124**, 1169–1181 (2006).
16. el-Deiry, W.S. *et al.* WAF1, a potential mediator of p53 tumor suppression. *Cell* **75**, 817–825 (1993).
17. Kim, T.K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
18. Léveillé, N. *et al.* Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA. *Nat. Commun.* **6**, 6520 (2015).
19. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
20. Fullwood, M.J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
21. Arnold, A. & Papanikolaou, A. Cyclin D1 in breast cancer pathogenesis. *J. Clin. Oncol.* **23**, 4215–4224 (2005).
22. Messeguer, X. *et al.* PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics* **18**, 333–334 (2002).
23. Sigoillot, F.D. & King, R.W. Vigilance and validation: Keys to success in RNAi screening. *ACS Chem. Biol.* **6**, 47–60 (2011).
24. Kearns, N.A. *et al.* Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nat. Methods* **12**, 401–403 (2015).
25. Kleinstiver, B.P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).
26. Canver, M.C. *et al.* BCL11A enhancer dissection by Cas9-mediated *in situ* saturating mutagenesis. *Nature* **527**, 192–197 (2015).

ONLINE METHODS

Cell lines and chemical reagents. BJ-RAS^{G12V}, HEK293-T, MCF-7, T47D and MDA-MB-231 cells were cultured in DMEM medium (Gibco), supplemented with 10% FCS (Hyclone), and 1% penicillin/streptomycin (Gibco). For the estrogen-depletion experiment, MCF-7 cells were cultured in DMEM phenol red–free medium (Gibco) supplemented with charcoal stripped serum (Gibco). 17β-estradiol was obtained from Sigma. All cell lines were obtained from the American Type Culture Collection, and they have been tested for mycoplasma contamination.

CRISPR-enhancer library design. For the p53-bound enhancers screen, we first took the union of the results of five publicly available p53 ChIP-seq analyses to create a combined set of 4,237 genomic sites that were bound by p53 in at least one cell line (MCF-7, CAL51 or IMR90) and in response to at least one stress (Nutlin-3a, 5-FU, RITA or ionizing radiation)[27,28]. We then scanned these sites for occurrences of the p53-binding motif using the *p53scan* tool[29] and found that 2,626 sites contained strong matches. To increase the chance that the candidate sites were functional ones, we intersected them with genomic locations of predicted enhancers in six different cell lines. These predictions, which are based on various histone marks, were downloaded from the UCSC Genome Browser (Broad ChromHMM track). 764 sites with a strong match for the p53-binding motif overlapped a predicted enhancer in at least one cell line. Last, we identified the sites that could be targeted by a CRISPR-Cas9 sgRNA that cleaves that DNA within the p53 binding motif, taking into account that Cas9 endonuclease cuts the DNA 3 nt upstream of the PAM (NGG). (In cases that *p53scan* tool predicted an occurrence of the p53 motif that contained a spacer, we required that the cleavage would be out of the spacer). Overall, we designed 1,116 sgRNA vectors that target the p53-binding motif within 685 different genomic binding sites. For the ERα-bound enhancers, we took the top 2,000 ChIP-seq binding sites[12] and identified the ERα consensus motif (up to one mismatch) in 740 of them. 406 of these sites could be targeted by the CRISPR-Cas9 system. As a further step to narrow down the candidate list and focus the screen on active enhancers, we intersected these regions with eRNA expression measured by the same study using GRO-seq[12]. Overall, 73 enhancers met these three criteria: (i) ERα binding detected by ChIP-seq; (ii) ERα motif that could be targeted by CRISPR-Cas9; and (iii) bidirectional eRNA expression. These 73 enhancers were targeted by 97 sgRNAs that comprise our CRISPR-ERα-enhLib. A list containing custom sgRNAs designed for this study can be found in **Supplementary Table 7**.

Pooled library cloning. We used standard de-salted DNA oligonucleotides, synthesized and purchased from IDT (Integrated DNA Technologies), to construct sgRNA libraries for p53-bound enhancers (1,116 sgRNAs), ERα-bound enhancers (97 sgRNAs) and deCDKN1A (197 sgRNAs). Complementary single-stranded oligos were phosphorylated and annealed by combining 100 µM oligos, 1× T4 PNK Buffer, 1 mM ATP, 5 U T4 PNK and incubating the reaction at 37 °C/30 min, 95 °C/5 min followed by a ramp down to 25 °C at 5 °C/min. Annealed oligos were pooled into three independent replicates (pool #1, #2, #3), diluted at 1:1,000 in sterile water, and ligated to plasmid vector lentiCRISPRv2 (gift from Feng Zhang (Addgene plasmid #52961)) using the following parameters: 50 ng BsmBI (Fermentas) digested plasmid, 1 µl diluted oligo duplex, 1× Ligation Buffer (Roche), 5 U T4 DNA Ligase (Roche) incubated at RT/30 min. We did five independent ligation reactions per pool, and used them to transform highly competent *Escherichia coli* cells (EletroSHOX - Bioline, BIO-85038) according to the manufacturer's protocol. In order to assess the complexity of our libraries, we plated 1 µl of cell transformation mixture on Luria-Bertani agar plates containing ampicillin, incubated them overnight at 37 °C, and counted individual bacterial colonies after 16 h. At this point, we estimated that each individual sgRNA is covered >100×, ensuring that our libraries have high complexity and are suitable for pooled screening. Transformation mixtures were combined, grew in liquid LB until $OD_{600} = 0.8$ was reached, and plasmid DNA was harvested using Genopure Plasmid Maxi kit (Roche).

Lentivirus production, purification and transduction. To produce lentivirus, 4×10^6 HEK293T cells per pool were seeded in ten 100-mm dishes 1 d before transfection. For each dish, we diluted 15 µg of CRISPR-enhancer plasmid library, 3.5 µg of pVSV-G, 5 µg of pMDL RRE and 2.5 µg of pRSV-REV in 450 µl of 0.1× TE/H_2O, added 50 µl of $CaCl_2$ and incubated 5 min at RT. Plasmid DNA was precipitated by adding 500 µl 2× HBS to the solution while vortexing at full speed. The precipitate was added immediately to the plate and the cells were incubated for 14 h at 37 °C, after which the medium was refreshed. Lentivirus-containing supernatants were collected 60 h post-transfection, filtered through a 0.45 µm membrane (Milipore Steriflip HV/PVDF) and stored at −80 °C. All cell types and lentivirus batches tested were titrated in order to achieve a multiplicity of infection of 0.4–0.5. Cell lines were infected with lentivirus supernatants supplemented with 8 µg/ml polybrene (Sigma). At 24 h post-infection, medium was replaced and cells were selected with 2 µg/ml puromycin (Gibco). Antibiotic selection was stopped as soon as no surviving cells remained in the no-transduction control plate.

CRISPR-Cas9 screen for OIS in BJ-RAS^{G12V} cells. In both OIS screens, we infected ~3,500 BJ-RAS^{G12V} cells per vector, to ensure that every sgRNA was present in the cell population at the start of the experiment. Cells infected with CRISPR-enhancer or nontargeting CRISPR library pools were allowed to proliferate for 48 h after antibiotic selection to clear potentially toxic sgRNAs from the population. At this time point, half of the cells infected with CRISPR-enhancer library pools were harvested ($T = 0$), and the remaining cells were placed in culture and treated with 100 nM tamoxifen (4-hydroxytamoxifen, Sigma) to induce HRAS^{G12V} expression. Cells infected with CRISPR-enhancer and nontargeting CRISPR pools were allowed to proliferate, while we monitored them for senescence or continuous proliferation. After 4 weeks of treatment, we harvested cells infected with CRISPR-enhancer pools ($T = 4$ weeks). Cell pellets harvested at $T = 0$ and $T = 4$ weeks were stored at −80 °C, and processed later on for further analysis. The validation of individual hits identified in both screens was done following the same procedures described above. Enrichment scores were calculated by comparing the normalized frequency of each sgRNA vector present in the cell populations at $T = 0$ with $T = 4$ weeks.

CRISPR-Cas9 dropout screen in breast cancer cells. MCF-7, T47D and MDA-MB-231 cells were infected with two independent pools of CRISPR-ERα-enhLib. We infected ~3,500 cells per vector, to ensure that every sgRNA was present in the cell population at the start of the experiment. Following antibiotic selection, cells were allowed to proliferate for 48 h to clear potentially toxic sgRNAs from the population. At this time point, we harvested half of the cells infected with CRISPR-ERα-enhLib pools ($T = 0$). The remaining cells were placed in culture, allowed to proliferate for 20 d, and then harvested ($T = 20$). Cell pellets were stored at −80 °C, and processed later on for further analysis. Enrichment (depletion) scores were calculated for each sgRNA vector, in each cell line (MCF-7, T47D and MDA-MB-231) by comparing its normalized frequency at $T = 20$ and $T = 0$ pools. Then the differences between these enrichment scores (in $\log_2$) were calculated for MCF-7 and T47D compared to MDA-MB-231 cells (which serve as controls in the screen as they are not dependent on ERα). These "Delta enrichment scores" calculated for MCF-7 and T47D were averaged and standardized (Z-scores). For validation, we selected sgRNA vectors whose repressive effect on proliferation was at least 30% stronger in both MCF-7 and T47D cells compared to MDA-MB-231 cells.

Genomic DNA sequencing to identify sgRNAs. Frozen cell pellets were thawed and genomic DNA (gDNA) was isolated with DNeasy Blood and Tissue kit (Qiagen). Identification of sgRNAs was done by PCR in two steps. For the first PCR, the amount of input gDNA was calculated to achieve >200× coverage over the CRISPR-enhancer libraries (assuming that 10^6 cells contain 6.6 µg gDNA), which resulted in 2 µg for CRISPR-p53-enhLib, 200 ng for CRISPR-ERα-enhLib and 300 ng for CRISPR-deCDKN1A-Lib. For each sample, we performed two separate reactions (max. 1 µg gDNA per reaction) using Phusion DNA polymerase (Thermo Scientific) and combined the resulting amplicons. In the first PCR, we used the following primer sequences to amplify lentiCRISPR-enhancer sgRNAs:

PCR1_F1

ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXXXGGCTTTATATATCTTGTGGAAAGGACG (XXXXXX represents a 6-bp barcode)

PCR1_R1

GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACTGACGGGCACCGGAGCCAATTCC

A second PCR was performed to attach Illumina adaptors and index samples. The second PCR was done in 50-µl reaction volume, including 5 µl of the product from the first PCR, and using the following primers:

PCR2_P5

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

PCR2_P7

CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT (XXXXXX represents a 6-bp index).

Amplification was carried out with 18 cycles for both first and second PCR. After the second PCR, resulting amplicons were purified using Agencourt AMPure XP beads (Beckman Coulter), quantified in a Bioanalyzer 2100 (Agilent), mixed and sequenced in a HiSeq 2500 (Illumina).

Identification of sgRNAs enriched in BJ-RASG12V cells. Based on deep sequencing of sgRNAs in BJ control and HRAS^{G12V}–induced populations (done in independent triplicates for each), we estimated the enrichment of each sgRNA vector in HRAS^{G12V} relative to control cells. This was done, by counting the number of reads corresponding to each sgRNA in each population, normalizing these counts to 1 M and taking the ratio of the normalized counts between HRAS^{G12V} and control cells. We averaged these enrichment scores (in $\log_2$ scale) for each sgRNA over the triplicates, and, last, calculate *Z*-scores (**Fig. 1c** and **Supplementary Table 1**). (To avoid inflation of ratios calculated for sgRNAs with low read counts, counts below 20 were set to 20.)

Genomic DNA sequencing to identify CRISPR-induced mutations. Cell pellets were collected and gDNA was isolated with DNeasy Blood and Tissue kit (Qiagen). Amplification of target regions for sequencing was done by PCR in two steps. For each sample, we used 500 ng of gDNA as input for the first PCR (done in duplicate). Resulting amplicons were combined and we used 5 µl as input for the second PCR. Amplification was carried out with 18 cycles for both first and second PCR. After the second PCR, amplicons were purified using Agencourt AMPure XP beads (Beckman Coulter), quantified in a Bioanalyzer 2100 (Agilent), mixed and sequenced in a HiSeq 2500 (Illumina).

Competitive proliferation assay. MCF-7 and MDA-MB-231 cells were infected with indicated sgRNAs to validate the results of the CRISPR-ERα-enhLib screen. Separately, we generated polyclonal MCF-7 and MDA-MB-231 cells stably expressing GFP using pLX304-GFP[30] (gift from David Root; Addgene plasmid # 25890). GFP expressing cells were mixed in a 1:3 ratio with cells containing individual sgRNAs. The percentage of GFP-expressing cells was assessed by flow cytometry at the beginning of the experiment ($T = 0$) and every 72 h onwards ($T = 3$ d and $T = 6$ d). For every condition, 10,000 events were recorded, and the data were analyzed using FlowJo software.

Western blot analysis. Whole-cell lysates were prepared as previously described[31]. Membranes were immunoblotted with the following antibodies: TP53 (DO-1, Santa Cruz; 1:1,000), CDKN1A (Sc-397, Santa Cruz; 1:1,000), HRAS (C-20, Santa Cruz; 1:1,000), Cyclin D1 (M-20, Santa Cruz; 1:1,000), HSP90 (H-114, Santa Cruz; 1:10,000), beta-Actin (C4, Santa Cruz, 1:10,000). Protein bands were visualized using corresponding secondary antibodies (Dako) and ECL reagent (GE Healthcare).

Senescence-associated β-gal assay. BJ-RAS^{G12V} cells were transduced with lentiCRISPRv2 constructs, selected with puromycin, plated in triplicate and treated for 15 d with 100 nM 4-OHT to induce HRAS^{G12V} expression. β-galactosidase activity was determined with Senescence β-galactosidase staining kit (Cell Signaling), and at least 100 cells were analyzed for each condition.

BrdU proliferation assays. Cells were pulsed for 3 h with 30 µM bromodeoxyuridine (BrdU, Sigma), fixed with ethanol (70% solution), permeabilized, treated with NaOH to denature DNA, incubated with anti-BrdU (GE Healthcare), washed in blocking buffer (PBS, Tween 0.05%, 2% BSA), and finally incubated with anti-rabbit AF488 secondary antibody (Dako). BrdU incorporation was measured either by immunofluorescence (at least 200 cells were scored for each condition) or by flow cytometry (10,000 events were recorded for each sample). Flow cytometry data were analyzed using FlowJo software.

Luciferase reporter assays. Sense and antisense region of *deCDKN1A* and *peCDKN1A* were PCR amplified from gDNA of BJ-RAS^{G12V} cells whereas ERα^{enh588} was amplified from MCF-7 cells. All regions were cloned into pGL3-promoter vector. Constructs were transfected into MCF-7 cells and treated with 8 µM Nutlin-3a (Cayman Chemical), 10^{-8} M 17β-estradiol (Sigma) or vehicle for 30 h. Reporter activity was measured 36 h after transfection using Dual-Luciferase system (Promega) according to the manufacturer's instructions.

RNA isolation, reverse-transcription and quantitative real-time PCR (qPCR). Total RNA was extracted using TRIsure (Bioline) reagent and following the manufacturer's protocol. cDNA was produced with SuperScript III (Invitrogen) using 5 µg of total RNA per reaction. qPCR reaction was performed with SYBR green I Master mix in a LightCycler 480 (Roche). TATA-binding protein (TBP) was used as an internal control. Primers used in qPCR are listed in **Supplementary Table 7**.

GRO-seq. GRO-seq was performed as described before with minor modifications. Briefly, 5×10^6 nuclei were isolated and incubated 5 min at 30 °C with equal volume of reaction buffer (10 mM Tris-Cl pH 8.0, 5 mM MgCl2, 1 mM DTT, 300 mM KCL, 20 units of SUPERase In, 1% sarkosyl, 500 µM ATP, GTP and Br-UTP, 0.2 µM CTP+32P CTP) for the nuclear run-on. The reaction was stopped and total RNA was extracted with Trizol LS (Invitrogen) according to the manufacturer's instructions. RNA was fragmented using fragmentation reagents (Ambion) and the reaction was purified through p-30 RNase-free spin column (BioRad). BrU-labeled RNA was immunoprecipitated with anti-BrdU agarose beads (Santa Cruz), washed one time in binding buffer, one time in low salt buffer (0.2× SSPE, 1 mM EDTA, 0.05% Tween-20), one time high-salt buffer (0.25× SSPE, 1 mM EDTA, 0.05% Tween-20, 137.5 mM NaCl) and two times in TET buffer (TE with 0.05% Tween-20). RNA was eluted with elution buffer (20 mM DTT, 300 mM NaCl, 5 mM Tris-Cl pH 7.5, 1 mM EDTA and 0.1% SDS) and isolated with Trizol LS. After the binding step, BrU-labeled RNA was treated with tobacco acid pyrophosphatase (TAP, Epicenter) to remove 5′-methyl guanosine cap, followed by T4 polynucleotide kinase (PNK; NEB) to remove 3′-phosphate group. BrU-containing RNA was treated with T4 PNK again at high pH in the presence of ATP to add 5′-phosphate group. The reaction was stopped and RNA was extracted with Trizol LS. Sequencing libraries were prepared using TruSeq Small RNA kit (Illumina) following manufacturers instructions. Briefly, end-repaired RNA was ligated to RNA 3′ and 5′ adapters, followed by RT-PCR amplification. cDNA was purified using Agencourt AMPure XP (Beckman Coulter) and amplified by PCR for 12 cycles. Finally, amplicons were cleaned and size-selected using Agencourt AMPure XP (Beckman Coulter), quantified in a Bioanalyzer 2100 (Agilent), and sequenced in a HiSeq 2500 (Illumina). Sequenced reads were aligned to the human genome (hg19) using bowtie2 (ref. 32).

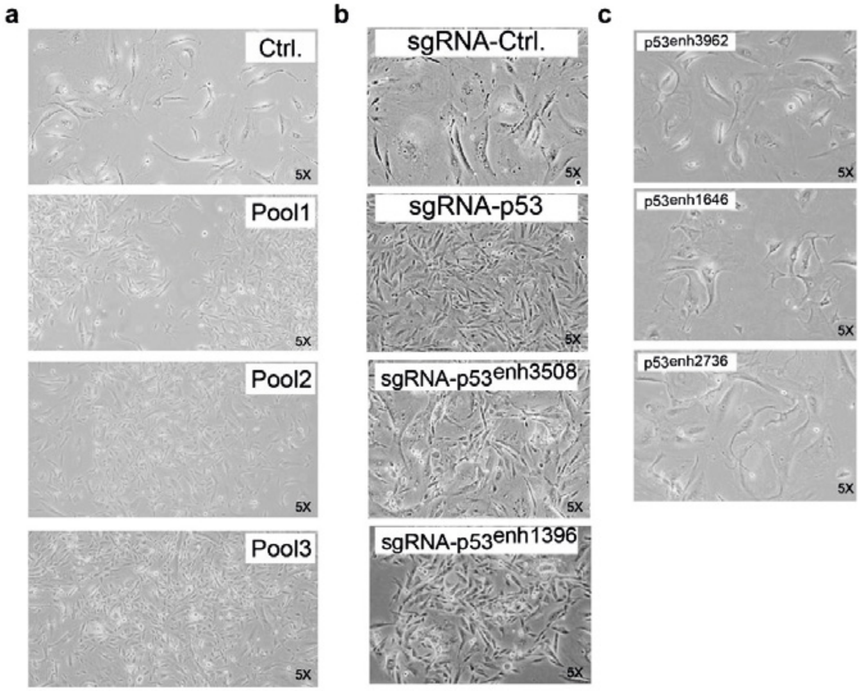
RNA-seq. RNA-seq samples were processed with TruSeq RNA library prep kit v2 (Illumina) and sequenced in a HiSeq 2500 (Illumina). Sequenced reads were aligned to the human genome (hg19) using TopHat2 (ref. 33) and gene expression counts were calculated using HTseq[34] based on Ensembl's human gene annotations (v69)[35]. Expression levels were normalized using quantile normalization.

ChIP. BJ-RAS^{G12V} (5×10^6) cells were fixed with 1% formaldehyde at RT/8min and quenched with 125 mM glycine for 5 min on ice. The cells were centrifuged at 470*g*/10 min and resuspended in 300 ml of cold lysis buffer (50 mM Tris-HCl, pH 8.0, 10 mM EDTA and 1% SDS) supplemented with protease inhibitor cocktail (Roche). The suspension was sonicated for 20 min (30 s on/off at maximum power) and diluted with 800 ml of dilution buffer (10 mM Tris-HCl, pH 7.5, 140 mM NaCl, 1 mM EDTA, 0.5 mM EGTA and 1% Triton X-100). The lysate was centrifuged at 14,000 r.p.m./10 min and the soluble fraction was transferred to a new tube. For each reaction, 100 ml of chromatin preparation was diluted in 300 ml of dilution buffer and incubated overnight with indicated antibody amount at 4 °C on a rotator. To each ChIP reaction, 30 ml of protein A/G beads, previously blocked (PBS/BSA (0.1%) for 1 h), was added and incubated 3 h at 4 °C. The immuno-precipitated chromatin was washed 2 × 5 min with dilution buffer and 1 × 5 min with TE (50 mM Tris-HCl pH 8.0 and 10 mM EDTA) and eluted overnight in 300 ml elution buffer (20 mM Tris-HCl pH 7.5, 5 mM EDTA, 50 mM NaCl and 1% SDS) at 65 °C in an orbital shaker. Eluted samples were purified using QIAquick PCR purification kit (Qiagen) and analyzed by real-time qPCR. The following antibodies and amounts were used in this experiment: 3 μg p53 (DO-1, Santa Cruz), 3 μg C/EBP beta (C-19, Santa Cruz). Primers used in ChIP-qPCR are listed in **Supplementary Table 7**.

ChIP-sequencing data analysis. Sequencing reads were aligned to the human genome (hg19) using bwa v 0.7.5 with default parameters. Number of aligned reads per sample can be found in **Supplementary Table 5**. Peaks in control MCF-7 and T47D cell lines were called with MACS[36] (default parameters) and DFilter[37] (parameters: –bs = 50 –ks = 30 –refine –nonzero) algorithms. In-house MCF-7 mixed input was used for peak calling of MCF-7 cell line; T47D input from a previous study[38] was used for calling T47D data. Intersect of the two peak calling algorithms was used for further analysis. 27020 and 6702 ERα peaks were detected in MCF-7 and T47D cell lines, respectively.

27. Botcheva, K., McCorkle, S.R., McCombie, W.R., Dunn, J.J. & Anderson, C.W. Distinct p53 genomic binding patterns in normal and cancer-derived human cells. *Cell Cycle* **10**, 4237–4249 (2011).
28. Rashi-Elkeles, S. *et al.* Parallel profiling of the transcriptome, cistrome, and epigenome in the cellular response to ionizing radiation. *Sci. Signal.* **7**, rs3 (2014).
29. Smeenk, L. *et al.* Characterization of genome-wide p53-binding sites upon stress response. *Nucleic Acids Res.* **36**, 3639–3654 (2008).
30. Yang, X. *et al.* A public genome-scale lentiviral expression library of human ORFs. *Nat. Methods* **8**, 659–661 (2011).
31. Agami, R. & Bernards, R. Distinct initiation and maintenance mechanisms cooperate to induce G1 cell cycle arrest in response to DNA damage. *Cell* **102**, 55–66 (2000).
32. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
33. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
34. Anders, S., Pyl, P.T. & Huber, W. HTseq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
35. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43**, D662–D669 (2015).
36. Zhang, Y. *et al.* Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008).
37. Kumar, V. *et al.* Uniform, optimal signal processing of mapped deep-sequencing data. *Nat. Biotechnol.* **31**, 615–622 (2013).
38. Ross-Innes, C.S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393. (2012).

Supplementary Figure 1

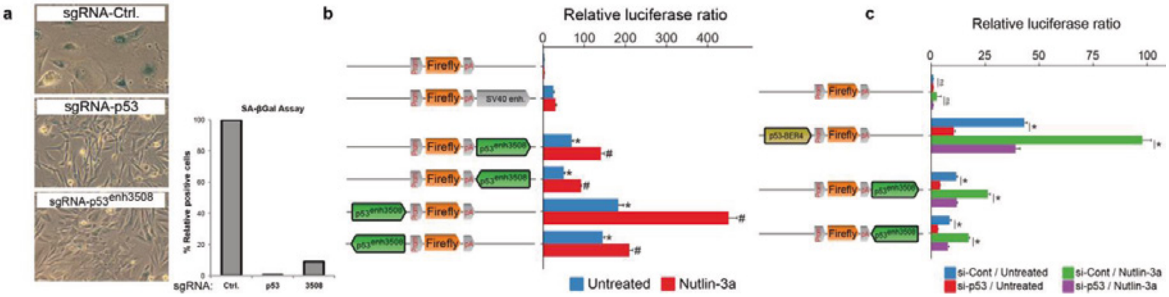


Supplementary Figure 1

Light microscopy images of cell populations transduced with the indicated sgRNA vectors.

(a, b and c) Images were taken after 15 days of HRAS^{G12V} induction.

Supplementary Figure 2

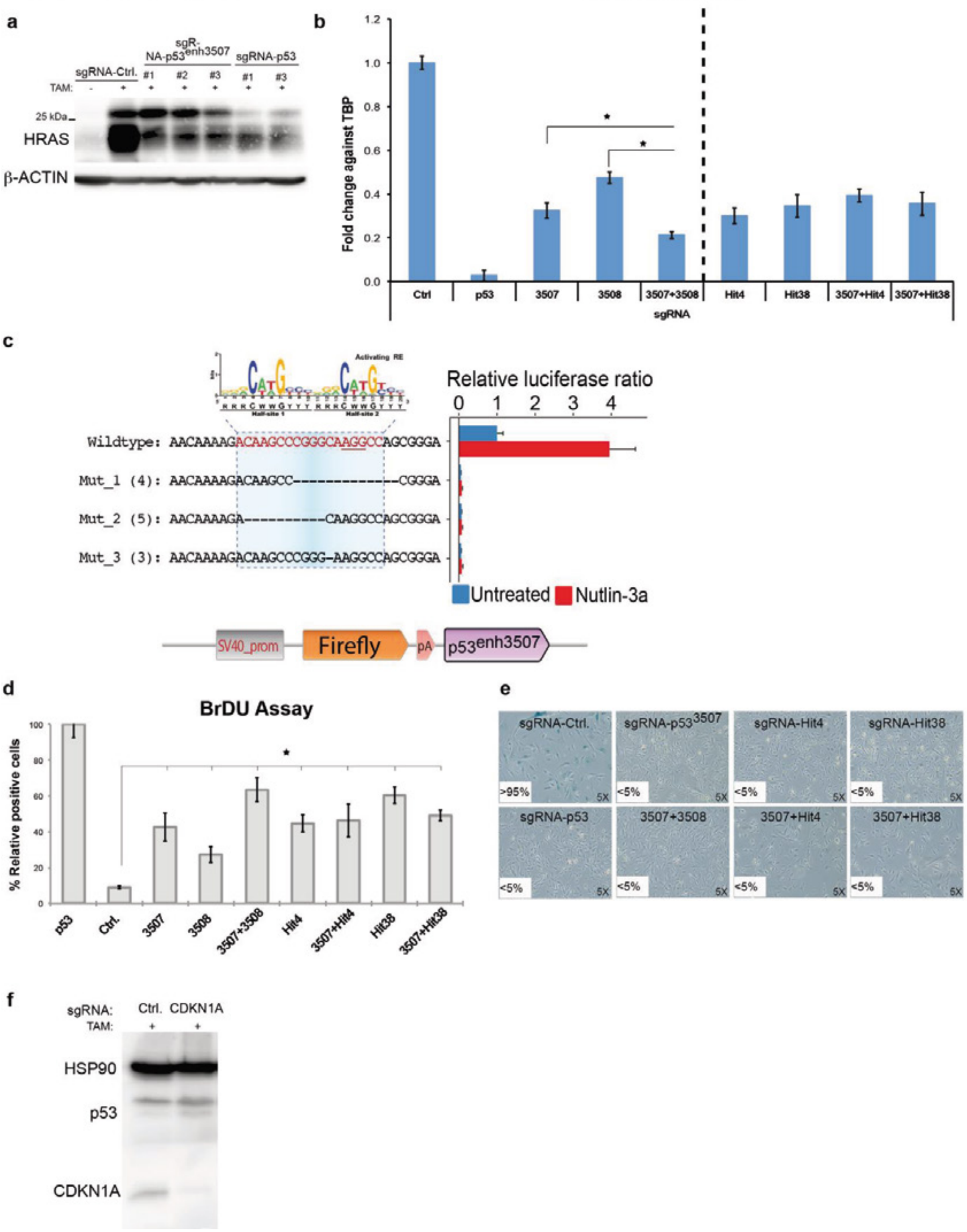


Supplementary Figure 2

p53^{enh3508} region exhibits a p53-dependent enhancer activity that is required for OIS activation.

(a) Senescence induction was quantified using senescence-associated β-gal assay. **(b)** MCF-7 cells were transfected with the indicated vectors, treated with Nutlin-3a 5-10 hours later, and harvested 25-30hr after treatment. The relative luciferase activities (Firefly/Renilla) were normalized to the Ctrl reaction. (All p-values for luciferase assay were calculated by Student's t-test. * indicates the significance relative to empty vector; # indicates the significance relative to untreated matching sample). **(c)** The same assay as in panel b, only that cells were co-transfected with control, or p53-targeting siRNAs. A reporter vector containing the enhancer region p53-BER4[20] was used as a positive control for p53-dependency. The efficiency of p53 knockdown was determined by immunoblot analysis. (All p-values for luciferase assay were calculated by Student's t-test. * indicates the significance relative to empty vector).

Supplementary Figure 3

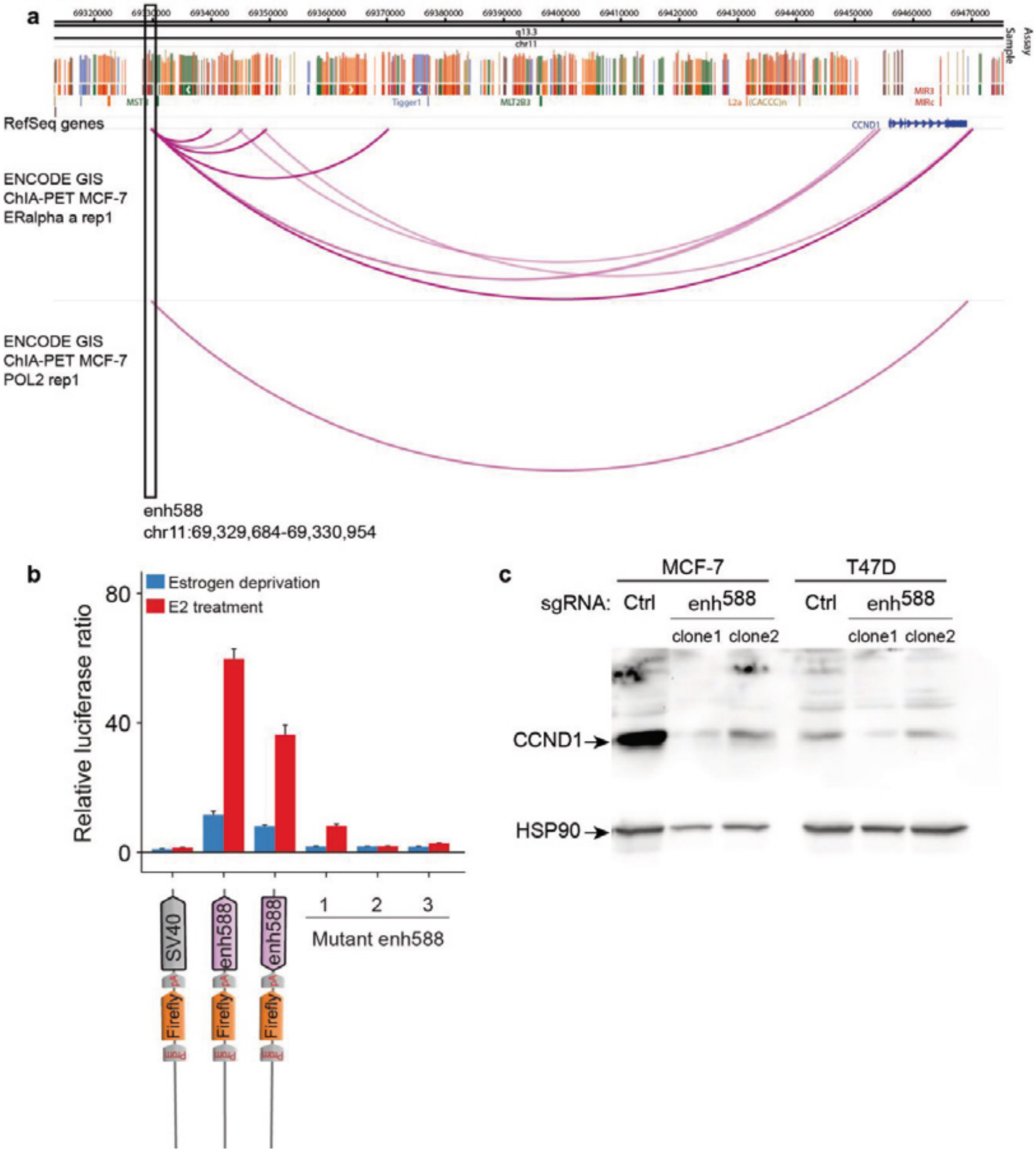


Supplementary Figure 3

deCDKN1A contains key enhancer elements that regulate OIS in a p53-dependet fashion.

(a) The same cell extracts as in Figure 2g were blotted with an antibody against HRAS. **(b)** qRT-PCR analysis of *CDKN1A* mRNA levels performed with the indicated BJ-indRAS^{G12V} cell populations following induction of RAS^{G12V}. (n=3; *P<0.05, two-tailed Student's t-test). **(c)** MCF-7 cells were transfected with the indicated reporter vectors, treated with Nutlin-3a 5-10 hours later, and harvested 25-30 hours later. The relative luciferase activities (Firefly/Renilla) were normalized to the Ctrl reaction. **(d and e)** The same cell populations as in b panel were subjected to BrdU labeling and β-gal assays to assess proliferation and OIS, respectively. N=2; for each condition at least 150 cells were count. *P<0.05, two-tailed Student's t-test. **(f)** Western blot analysis of BJ-indRAS^{G12V} CDKN1A KO and control cells. P53 and HSP90 were used as controls.

Supplementary Figure 4

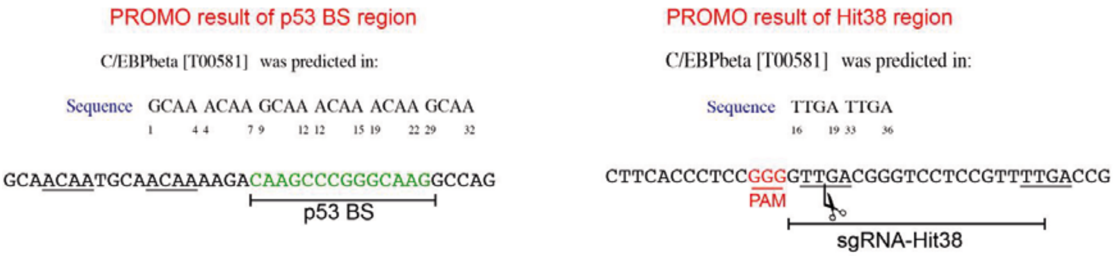


Supplementary Figure 4

ERα^{enh588} regulates *CCND1* activity in an estrogen-dependent manner.

(a) Screenshot of ChIA-PET analysis in MCF-7 cells showing a strong chromatin interaction between enh588 and *CCND1* promoter. **(b)** MCF-7 cells were seeded with charcoal-treated medium and transfected with the indicated reporter vectors. 16 hours later the medium was refreshed and either supplemented or not with 17β-estradiol. 24 hours later cells were harvested and luciferase activity was measured. The relative luciferase activities (Firefly/Renilla) were normalized to the Ctrl reaction. **(c)** Complete blot shown in Figure 3h.

Supplementary Figure 5


PROMO result of p53 BS region
C/EBPbeta [T00581] was predicted in:
Sequence GCAA ACAA GCAA ACAA ACAA GCAA
1 4 4 7 9 12 12 15 19 22 29 32
GCAACAATGCAACAAAAGACAAGCCCGGGCAAGGCCAG
p53 BS
PROMO result of Hit38 region
C/EBPbeta [T00581] was predicted in:
Sequence TTGA TTGA
16 19 33 36
CTTCACCCTCCGGGGTTGACGGTCCTCCGTTTTGACCG
PAM
sgRNA-Hit38

Supplementary Figure 5

Results of PROMO analyses are shown for regions surrounding the p53 BS and Hit38.

The p53 BS and the sgRNA-Hit38 targeting region are indicated, and the sequences representing potential CEBPB binding are underlined.

Chapter 5

CUEDC1 is a primary target of ERα that is essential for the growth of breast cancer cells

Manuscript in preparation

CUEDC1 is a primary target of ERα that is essential for the growth of breast cancer cells

Rui Lopes[1,4], Gözde Korkmaz[1,4], Sonia Aristín Revilla[1], Lars Custers[1], Yongsoo Kim[1], Arieh Tal[1], Pieter C. van Breugel[1], Wilbert Zwart[1], Ran Elkon[3], Reuven Agami[1,2,5].

[1]Division of Oncogenomics, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands
[2]Department of Molecular Genetics, Erasmus University Medical Center, Wytemaweg 80, 3015 CN Rotterdam, the Netherlands
[3]Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel
[4]These authors contributed equally
[5]Corresponding author

Abstract

Breast cancer is the most prevalent type of malignancy in women, and each year ~1.7 million new cases are diagnosed. Nearly 70% of breast tumors express *ERα* (*ESR1*), which is a ligand-dependent transcription factor (TF) that activates the expression of critical genes involved in cell proliferation (e.g. *CCND1* and *MYC*). The main treatment of ERα-positive breast cancer is based on hormonal therapies (e.g. tamoxifen and aromatase inhibitors), which aim to inhibit the activity of ERα in cancer cells. Despite the extensive use of these drugs, a substantial number of tumors relapse after initial treatments and eventually patients develop resistance to therapy. Importantly, ChIP-seq experiments revealed that the vast majority of ERα binding events map to regions that have features of enhancer elements. Moreover, it was found that ERα controls the expression of ~3,000 genes and ~1200 enhancers in breast cancer cells. This evidence underlines the great challenge of identifying the direct target genes of ERα and understanding their contribution to the proliferation of cancer cells. Recently, we performed genetic screens to identify enhancers that are required for the growth of ERα-positive breast cancer cells. We validated several candidates, including a putative enhancer located in the first intron of *CUEDC1*. Here, we show that CUTE (CUEDC1 Transcriptional Enhancer) is a ERα-responsive enhancer that controls the expression of *CUEDC1* in breast cancer cells. Moreover, genetic alterations in CUTE decrease the expression of *CUEDC1* and reduce the proliferation of cancer cells. Finally, the expression of *CUEDC1* is increased in ERα-positive tumors and this is associated with poor clinical outcome of cancer patients. Altogether, our work suggests that *CUEDC1* is a primary target gene of ERα and a potential biomarker of human breast cancer.

Introduction

Breast cancer is the most prevalent type of malignancy in women, and each year ~1.7 million new cases are diagnosed[1]. Approximately 70% of breast tumors express *ERα* (*ESR1*), which is a ligand-dependent transcription factor (TF) that plays a critical role in cell proliferation[2]. ERα is activated by estradiol (E2), which is its natural ligand, or through phosphorylation events mediated by kinases such as MAPK/PI3K[2]. ERα is a ligand-dependent TF that is recruited directly or indirectly to the chromatin, in order to activate the expression of its target genes[3]. The current treatment of ERα-positive breast cancer is mainly based on hormonal therapies, which either compete with E2 for binding to ERα (e.g. tamoxifen)[4] or prevent the synthesis of E2 (e.g. aromatase inhibitors)[5]. Despite the extensive use of these drugs, a substantial number of tumors relapse after initial treatments and patients develop resistance to therapy[6].

ERα regulates the expression of several genes that play a central role in the development of breast cancer, including *CCND1*, *E2F1* and *MYC*[7]. ChIP-seq experiments revealed that the vast majority of ERα binding events map to regions that are distantly located from gene promoters and have features of enhancer elements[8,9]. Enhancers are non-coding regulatory elements that control gene expression in space and time, which is essential for specifying different cell-lineages during the development of organisms[10]. Therefore, it is not surprising that genetic alterations in enhancers are associated with the development of cancer[11,12]. Genome-wide mapping of nascent RNA by GRO-seq identified ~1200 enhancers that express eRNAs upon activation of ERα in breast cancer cells[13,14]. eRNA expression is a hallmark of active enhancer elements[15-18] and there is abundant evidence suggesting that these transcripts are required for the activation of target genes of ERα[13,14]. The studies mentioned above contributed to elucidate the mode of action of ERα and provided a comprehensive map of its binding sites throughout the genome[8,13,14]. However, they are inherently descriptive and do not fully explain the function and mechanisms of ERα-regulated enhancers. Moreover, it is not clear which are the primary target genes of ERα and how they contribute to the proliferation of breast cancer cells. Therefore, answering these questions is critical to elucidate the role of ERα in gene regulation, and to improve current diagnosis and therapies of breast cancer.

Traditionally, it is challenging to study the function of enhancers due to a lack of genetic tools to manipulate them in a high-throughput manner. The development of CRISPR-Cas9 systems opened exciting possibilities for targeted genome editing. Of note, Cas9 can be directed to virtually any genomic sequence by a single guide RNA (sgRNA), provided that there is a protospacer adjacent motif (PAM) downstream of the target sequence[19]. Cas9 is a nuclease that efficiently induces double-strand breaks (DSBs), which give rise to small insertions and deletions when repaired by non-homologous end joining (NHEJ). Cas9 is capable of cleaving multiple target sequences in parallel[20,21], making it particularly suitable to perform genome-wide genetic screens[22,23]. Recently, we and others pioneered the application of CRISPR-Cas9 to map functional regulatory elements in human cells[24-27]. In our work, we performed genetic screens to identify enhancers that are required for the growth of ERα-positive breast cancer cells[25]. We validated several candidate hits from the screen, including a putative enhancer located in the first intron of *CUEDC1* (CUE Domain Containing 1) – a poorly characterized gene that was not previously associated with breast cancer. Here, we show that CUTE (CUEDC1 Transcriptional Enhancer) is a *bona fide* enhancer that activates *CUEDC1* expression in response to ERα signaling. The inactivation of CUTE by genetic alterations decreases the expression of *CUEDC1* and reduces the proliferation of cancer cells. Finally, we found that the expression of *CUEDC1* is significantly increased in ERα-positive tumors and this is associated with poor clinical outcome of cancer patients. Altogether, our work suggests that *CUEDC1* is a primary target gene of ERα that is required for the growth of breast cancer cells *in vivo*.

Results

To study the function of CUTE, we started by analyzing the genomic landscape around its locus. CUTE is located in the first intron of *CUEDC1* (**Fig. 1A**; yellow vertical line), and this region is predicted to be an enhancer in different human cell-types according to ENCODE data (**Fig. 1B**; yellow and orange colors). In MCF7 cells (ERα-positive), the CUTE locus is marked by high H3K27Ac and low H3K4me3 levels (**Fig. 1C**). This pattern of histone modifications is indicative of enhancers and often used to annotate this type of elements[28,29]. The H3K27Ac signal spans over 8 kb and encompasses three DNase I-hypersensitive sites (DHSs) (**Fig. 1C**). Of note, the DNase I signal overlapping with CUTE is increased upon treatment with E2 (**Fig. 1C**), suggesting that it is an ERα-responsive regulatory element. In line with this, we detected transcription of eRNAs by GRO-seq in MCF7 cells (**Fig. 1D**). Importantly, the expression of eRNAs in MCF7 cells is increased upon activation of ERα[14], whereas in MDA-MB-231 (ERα-negative) they are not detectable (**Fig. 1D**). Finally, we observed that the binding of ERα at CUTE overlaps with p300 (**Fig. 1E**), which is a coactivator protein that is often found at enhancers[30]. Altogether, these results suggest that CUTE is a putative enhancer regulated by ERα in breast cancer cells.

Enhancers are known to activate gene expression regardless of their orientation relative to the target gene[31]. To address this point, we cloned a DNA fragment containing CUTE (~1 kb), in forward and reverse orientations, into pGL3-promoter vector. We transfected these constructs into MCF7 cells and observed a strong activation of luciferase activity (~8 fold; *P*-value<0.001) independently of the orientation of CUTE (**Fig. 2A**). Next, we tested the responsiveness of CUTE to estrogen activation by transfecting the constructs into MCF7 cells and treating them with E2. We observed a significant increase (~30 fold; *P*-value<0.001) in the activity of luciferase (**Fig. 2B**), which indicates that the enhancing capacity of CUTE is dependent of ERα. To test this hypothesis, we generated mutations in the estrogen-responsive element

(ERE) of CUTE using CRISPR-Cas9 gene editing (**Fig. 2C**). We observed that the transcriptional activity of CUTE is severely compromised by small deletions (<10 bp) in its ERE (**Fig. 2B,C**), thereby confirming that CUTE is a *bona fide* enhancer regulated by ERα.

As mentioned above, we identified CUTE as a putative enhancer in a previous CRISPR-Cas9 screen[25] (candidate sg1830). Here, we transduced MCF7 cells with sg1830, which targets the ERE of CUTE (**Fig. 3A**), and performed targeted DNA-seq of this region to assess the effects of CRISPR-Cas9 editing (**Fig. 3B**). We observed that the vast majority of deletions generated by Cas9 are small (<5 bp) and map to the expected cleavage site (**Fig. 3B**). Next, we measured the binding of ERα by ChIP-seq in WT (sgCtrl) and mutant (sg1830) MCF7 cells to test the specificity of CRISPR-Cas9 targeting. Remarkably, the binding of ERα is only decreased at the CUTE locus (e1830) in mutant MCF7 cells (Pearson = 0.92; *P*-value = 0) (**Fig. 3C**). Of note, two known EREs (e588 and e1896)[25] were not affected by sg1830 (**Fig. 3C**). It is known that the binding of ERα is frequently associated with FOXA1, and this is thought to be required for the activation of target genes of estrogen[32]. Interestingly, FOXA1 binding to CUTE was also decreased in MCF7 cells expressing sg1830 (Pearson = 0.94; *P*-value = 0) (**Fig. 3D**). These results suggest that the disruption of ERα binding causes loss of FOXA1, which can affect the enhancing activity of CUTE. Consistent with this hypothesis, we found that the activating histone mark H3K27Ac is significantly decreased at the CUTE locus in mutant MCF7 cells (Pearson = 0.96; *P*-value = 0) (**Fig. 3E**). Our results indicate that sg1830 specifically impairs the binding of ERα to CUTE, which may compromise the function of this enhancer.

The identification of functional enhancers and their target genes is a major challenge in the field of transcriptional research[33]. Here, we combine multiple techniques and analyzes to identify the target gene of CUTE in breast cancer cells. First, we reanalyzed chromatin interactions identified by ChIA-PET and found that CUTE interacts with several regions nearby *CUEDC1* (**Fig. 4A**). Interestingly, the interactions with the promoter of *CUEDC1* involve both ERα and RNAPII (**Fig. 4A**), suggesting that they might be functional. Then, we performed RNA-seq in WT and mutant MCF7 cells in order to identify genes regulated by CUTE. We analyzed changes in gene expression in the vicinity of CUTE (+/- 500 kb) and found that *CUEDC1* is the most downregulated gene in mutant MCF7 cells (**Fig. 4B**). We confirmed by qPCR analysis that the expression of *CUEDC1* is significantly decreased in mutant MCF7 cells (~70%; *P*-value<0.01) (**Fig. 4C**). On the other hand, *CUEDC1* expression was not affected by sg1830 in MDA-MB-231 cells (**Fig. 4C**), suggesting that CUTE is active in ERα-positive cells. In line with this, the expression CUTE eRNAs is decreased in mutant MCF7 but not in mutant MDA-MB-231 cells (**Fig. 4D**). Next, we reanalyzed publicly available GRO-seq data of MCF7 cells treated with E2 in order to identify genes regulated by ERα. We focused our analysis on the CUTE locus (+/- 500 kb), and identified *CUEDC1* as the most upregulated gene upon ERα activation (**Fig. 4E**). We validated these results by performing a time-course experiment in MCF7 cells treated with E2. We found that the expression of *CUEDC1* is significantly increased (~2.5 fold) already 4h after E2 treatment and remains relatively stable over the course of 24 h (**Fig. 4F**), indicating that *CUEDC1* is a primary target gene of ERα. The expression of eRNAs transcribed from CUTE follow a similar pattern in MCF7 cells (**Fig. 4F**), supporting the notion that CUTE is responsive to ERα. So far, our results suggest that CUTE mediates the activation of *CUEDC1* expression through ERα. In order to test this hypothesis, we treated WT and mutant MCF7 cells with E2 and measured the expression of *CUEDC1* by qPCR. Reassuringly, we found that the induction of *CUEDC1* expression by E2 was severely compromised in mutant MCF7 cells (**Fig. 4G**), thereby confirming that CUTE is an enhancer of *CUEDC1* in breast cancer cells.

We showed previously that the mutation of CUTE by CRISPR-Cas9 (sg1830) is associated with decreased growth of MCF7 cells[25]. Given that CUTE is an enhancer of *CUEDC1*, we hypothesized that this gene is required for cell growth mediated by ERα. Indeed, we observed a significant decrease in the proliferation of MCF7 cells transduced with two different sgRNAs targeting the coding sequence of *CUEDC1* (sgCUEDC1 exon#1 and exon#2) (**Fig. 5A**). These effects seem to be specific since two sgRNAs targeting a different intron of *CUEDC1* (sgCUEDC1 intron#1 and intron#2) did not cause substantial effects on cell growth (**Fig. 5A**). Of note, the sgRNAs targeting *CUEDC1* (sgCUEDC1 exon#1 and exon#2) phenocopy the disruption of CUTE (sg1830) (**Fig. 5A**), suggesting that this gene is required for the growth of MCF7 cells. Indeed, we found that the ectopic expression of *CUEDC1* in mutant MCF7 cells completely rescues the proliferative defect caused by the mutation of CUTE (**Fig. 5B,C**).

Our results demonstrate that *CUEDC1* is a target gene of ERα that is essential for the proliferation of MCF7 cells. If this holds true *in vivo*, the expression of *CUEDC1* should be associated in a positive manner with *ERα* expression in breast tumors. To test this hypothesis, we reanalyzed breast cancer samples (n=759) from TCGA and found that *CUEDC1* expression is significantly higher in ERα-positive than in ERα-negative tumors (*P*-value = 10^{-10}) (**Fig. 6A**). Moreover, the expression of *CUEDC1* is positively correlated with ERα expression in breast tumors (R = 0.3; *P*-value = 10^{-7}) (**Fig. 6B**). These results indicate that *CUEDC1* might be essential for the growth of breast cancer cells *in vivo*. Finally, we evaluated the prognostic value of *CUEDC1* expression for the clinical outcome of breast cancer patients. For this purpose, we reanalyzed publicly available data corresponding to 1,809 samples[34] and found that elevated *CUEDC1* expression is associated with decreased overall survival (HR = 1.04) (**Fig. 6C**) and distant metastasis-free survival (HR = 1.16) (**Fig. 6D**), although these trends are not statistically significant. Importantly, high expression of *CUEDC1* predicts worst relapse-free survival for patients (HR = 1.25; *P*-value = 0.0001) (**Fig. 6E**), suggesting that *CUEDC1* is a potential biomarker of resistance to breast cancer therapy.

Discussion

The identification of direct target genes of ERα is a fundamental question in current breast cancer research[35]. Genome-wide analysis using GRO-seq identified ~3,000 protein-coding genes that are regulated by estrogen in breast cancer cells[17]. This number is substantially higher than what was previously determined using expression microarrays, and corresponds to ~33% of all expressed genes in MCF7[17]. Despite these advances, it is not clear which are the direct targets of this pathway since ERα mostly binds to distal enhancer elements[8,9]. Recently, we showed that functional enhancers can be annotated in an unbiased manner by CRISPR-Cas9 screens[25,36]. Moreover, we demonstrated that combining genome editing with differential gene expression analysis is a powerful method to identify the target genes of enhancers. Here, we applied our strategy to characterize CUTE and found multiple evidence suggesting that it is an enhancer is responsive to estrogen and regulates the expression of *CUEDC1* in breast cancer cells: first, the binding of ERα to CUTE in MCF7 cells is increased upon treatment with E2 (**Fig. 1E**); second, mutations in the ERE of CUTE specifically decrease the binding of ERα in this region (**Fig. 3C**); third, long-range chromatin interactions between CUTE and the promoter of *CUEDC1* involve ERα (**Fig. 4A**); fourth, the activation of *CUEDC1* expression by estrogen is totally dependent on the ERE (**Fig. 4G**); finally, the expression of *CUEDC1* is positively correlated with ERα expression in breast tumors (**Fig. 6A,B**).

Our findings suggest that CRISPR-Cas9 has high specificity of targeting at enhancer elements. This is supported by ChIP-seq data showing that the binding of ERα is specifically decreased at CUTE by sg1830 (**Fig. 3C**). The fact that a single locus is affected by CRISPR-mediated gene editing is quite remarkable since ERα binds to tens of thousands of EREs across the genome[37]. We observed similar effects at additional ERα-regulated enhancers in multiple breast cancer cell lines[25], further supporting the application of CRISPR-Cas9 to characterize transcriptional enhancers *in situ*[36]. We observed that the loss of ERα binding is associated with decreased binding of FOXA1 and H3K27Ac at the CUTE locus in mutant MCF7 cells (**Fig. 3D,E**). FOXA1 is a pioneer factor that is required for ERα-mediated gene regulation[32], whereas H3K27Ac is a mark frequently associated with active enhancers[28,29]. Additionally, the expression of eRNAs is decreased in mutant MCF7 cells (**Fig. 4D**), suggesting that the activity of RNAPII is impaired. We conclude that ERα is an essential component of CUTE since the disruption of binding causes loss of enhancer-associated marks and decreased transcriptional activity. However, we cannot rule out that there are additional sequences, besides the ERα binding site, which are critical for the activity of CUTE. This question can be addressed by saturation mutagenesis experiments using CRISPR-Cas9, which allow dissecting regulatory DNA sequences at near-nucleotide resolution[24,25,27].

The concept of “super-enhancers” has been recently proposed to describe regulatory elements that drive exceptionally high levels of transcription[38-40]. Super-enhancers typically comprise a cluster of regulatory elements, spanning up to 12.5 kb, which exhibit highly synergistic activities[41]. Our results suggest that CUTE is a strong enhancer in human cells: in reporter assays, it is able to activate gene expression in a robust manner (~30 fold upon ERα activation) (**Fig. 2B**); and genetic alterations in its sequence cause a dramatic reduction in *CUEDC1* expression (~70%) (**Fig. 4C**). Interestingly, we noted that the CUTE locus fulfils the bioinformatic criteria of super-enhancers[38]: extended signal of H3K27Ac (~8 kb), multiple DHSs located in close proximity (~4kb) and active transcription by RNAPII (**Fig. 1**).

Moreover, the two DHSs in the vicinity of CUTE are also bound by cell-type-specific TFs (e.g. FOXA1, GATA3; data not shown) and co-activator p300 (**Fig. 1E**). Given the available evidence, it is tempting to speculate that these DHSs are enhancers that collaborate with CUTE to regulate the expression of *CUEDC1*. However, further experiments (e.g. deletion of each constituent element) are required to clarify whether they are functional elements and if they act as a super-enhancer of gene expression[42,43].

To date, not much is known about the biological functions of *CUEDC1*, though available evidence points in the direction of the ubiquitin pathway. CUE domains are sequences of ~40 amino acids that bind monoubiquitin in yeast and human[44] and regulate chain formation by E3 ligases[45]. Interestingly, CUEDC1 was recently identified in a proteome-wide screen for ubiquitin interactors[46]. Zhang and colleagues showed that CUEDC1 binds to K33 and K63 diubiquitin in different human cell-types, suggesting that it might be involved in protein trafficking, signal transduction and degradation pathways[47]. The connection of *CUEDC1* to cancer is thinner, although its expression is significantly upregulated in metastatic cervical tumors compared to primary tumors[48]. In conclusion, our work revealed that *CUEDC1* is a direct target gene of ERα that is essential for the growth of cancer cells (**Fig. 5B**), and suggest that it is a potential biomarker of human breast cancer.

Figures

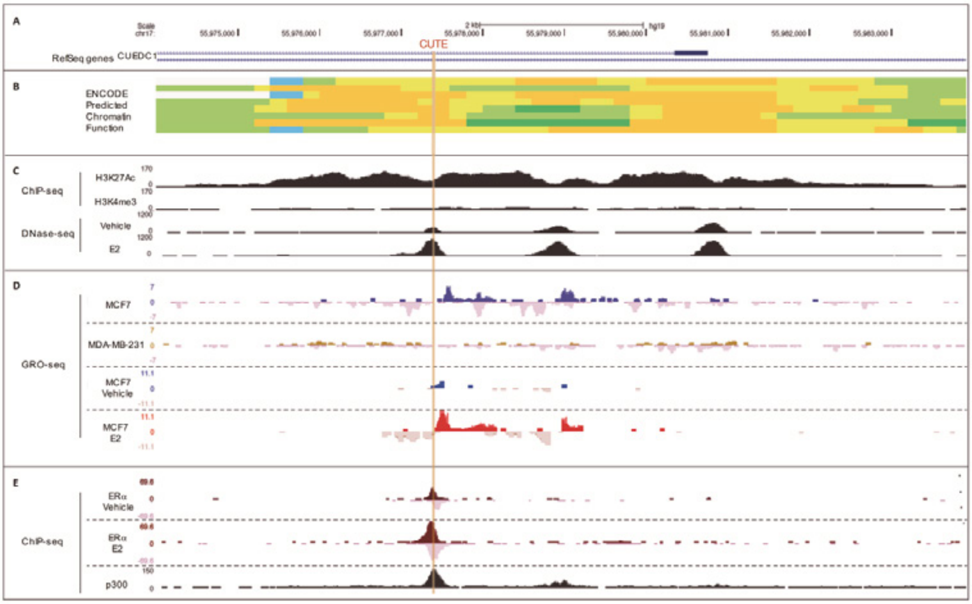


Figure 1 CUTE is a putative enhancer element regulated by ERα. (**A**) RefSeq genes track from NCBI showing the genomic location of CUTE (yellow vertical line). (**B**) Chromatin State Segmentation by a hidden Markov model from ENCODE/Broad in eight human cell-types. Color code: yellow/orange - enhancer; green - active transcription; blue - insulator; grey - low signal. (**C**) ChIP-seq (H3K27Ac and H3K4me3) and DNase-seq (Vehicle and E2) data of MCF7 cells. (**D**) GRO-seq data of MCF7 and MDA-MB-231 cells. The two bottom tracks display GRO-seq data of MCF7 cells treated with vehicle and E2[14]. Sequencing reads mapping to the sense and antisense DNA strands are displayed as positive and negative values, respectively. (**E**) ChIP-seq of ERα in MCF7 cells treated with vehicle and E2. The bottom track corresponds to ChIP-seq data of p300 in MCF7 cells.

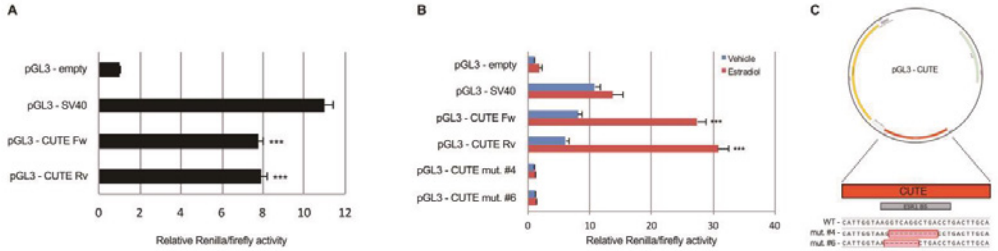


Figure 2 The transcriptional activity of CUTE is dependent of ERα. (A) MCF7 cells were co-transfected with *Renilla* and pGL3-based vectors and luciferase activities were measured after 48h. The SV40 enhancer (pGL3-SV40) was used as a positive control in this assay. The relative luciferase activities (*Renilla*/firefly) were normalized to pGL3-empty. Data represent mean ± s.d of n = 3. ****P* < 0.001, two-tailed Student's t-test relative to pGL3-empty. Fw, forward. RV, reverse. (B) MCF7 cells were co-transfected with *Renilla* and pGL3-based vectors and treated with vehicle or E2 for 24h. The relative luciferase activities (*Renilla*/firefly) were normalized to pGL3-empty vehicle. Data represent mean ± s.d of n = 3. ****P* < 0.001, two-tailed Student's t-test relative to pGL3-empty vehicle. (C) Schematic representation of pGL3-CUTE constructs used in reporter assays. The mutant sequences of CUTE (mut. #4 and #6) were generated by CRISPR-Cas9 gene editing in MCF7 cells, amplified by PCR and cloned into pGL3-promoter.

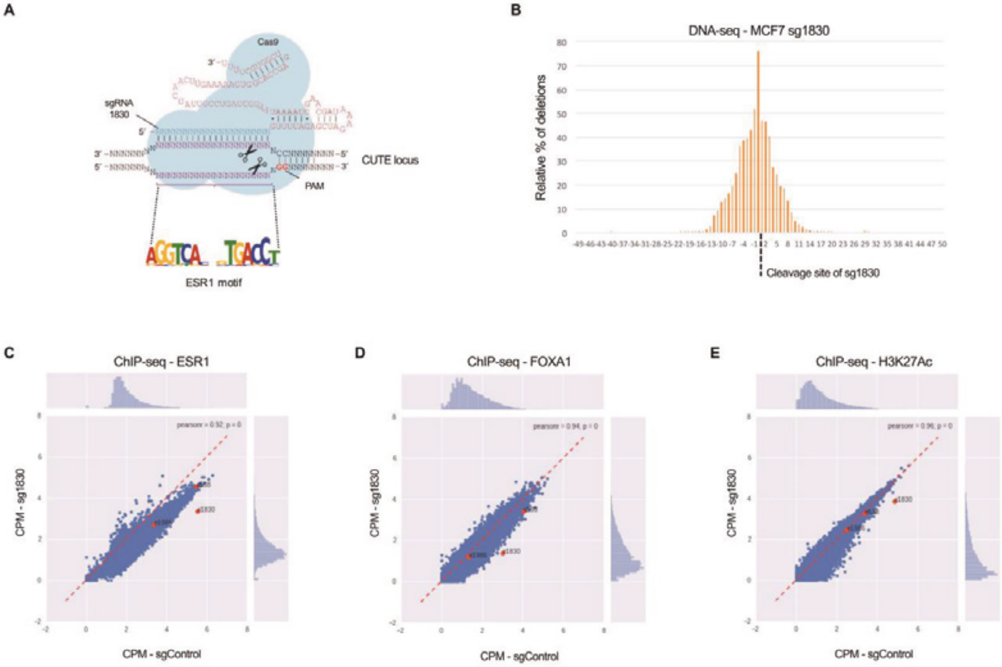


Figure 3 The specific disruption of ERα binding by sg1830 is associated with loss of FOXA1 and H3K27Ac binding at the CUTE locus. (A) Schematic representation of the targeting of CUTE by CRISPR-Cas9 (sg1830). (B) DNA-seq profile of the CUTE locus (100 bp) in MCF7 cells stably expressing sg1830. The prevalence of deletions that occurred at each position within this interval was calculated relative to the total number of reads that contained any deletion. (C-E) ChIP-seq of ERα (C), FOXA1 (D) and H3K27Ac (E) in WT (sgControl) and mutant (sg1830) MCF7 cells. e588 - ERE 588. e1830 - ERE 1830. e1986 - ERE 1986. CPM, counts per million.

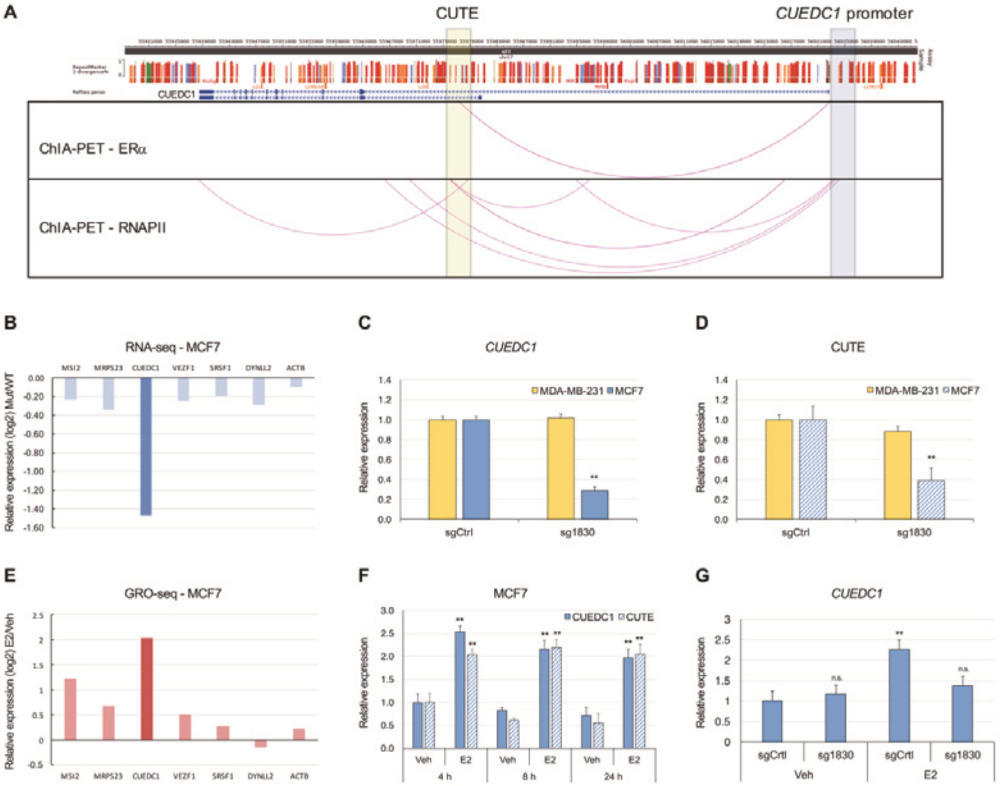


Figure 4 CUTE is an ERα-responsive enhancer that regulates the expression of *CUEDC1* (A) ChIA-PET data of ERα and RNAPII in MCF7 cells. Violet arcs represent long-range chromatin interactions that are statistically significant. (B) RNA-seq was performed in MCF7 cells transduced with sgCtrl (WT) and sg1830 (Mut). Differential expression analysis was restricted to genes located -/+ 500 Kb of CUTE. ACTB was used as a negative control. (C, D) Analysis of *CUEDC1* mRNA (C) and CUTE eRNA (D) expression by qPCR in WT (sgCtrl) and mutant (sg1830) MCF7 and MDA-MB-231 cells. Gene expression levels are normalized to *TBP*. Data represent mean ± s.d of n = 3. **P < 0.01, two-tailed Student's t-test relative to MCF7 sgCtrl (E) GRO-seq data of MCF7 cells treated with vehicle (Veh) and estradiol (E2). Differential expression analysis was restricted to genes located -/+ 500 Kb of CUTE. ACTB was used as a negative control. (F) Analysis of *CUEDC1* mRNA (solid bars) and CUTE eRNA (striped bars) expression by qPCR in MCF7 cells treated with vehicle (Veh) or estradiol (E2). Gene expression levels are normalized to *TBP*. Data represent mean ± s.d of n = 3. **P < 0.01., two-tailed Student's t-test relative to Vehicle-4 h (G) Analysis of *CUEDC1* mRNA expression by qPCR in WT (sgCtrl) and mutant (sg1830) MCF7 cells treated with vehicle (Veh) or estradiol (E2). *CUEDC1* expression levels are normalized to *TBP*. Data represent mean ± s.d of n = 3. **P < 0.01, two-tailed Student's t-test relative to sgCtrl-Vehicle. n.s., non-significant.

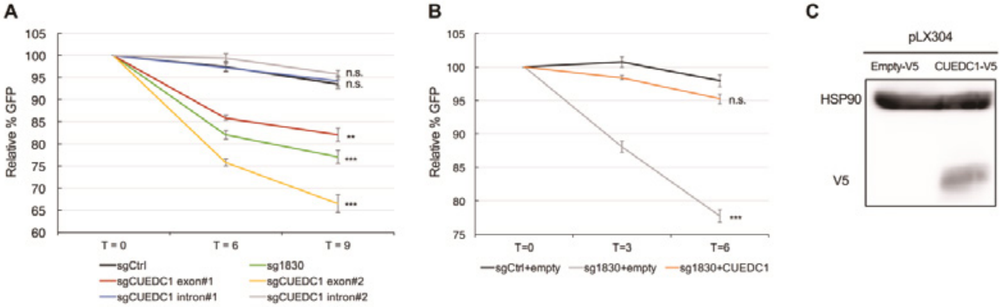


Figure 5 *CUEDC1* is required for the growth of MCF7 cells. (A) MCF7 cells were transduced with the indicated sgRNAs and allowed to proliferate for nine days. Cell growth is represented as percentage of GFP, which is normalized to the beginning of the experiment (T = 0). Data represent mean ± s.d of *n* = 3. ***P* < 0.01. ****P* < 0.001. n.s., non-significant. Two-tailed Student's t-test relative to sgCtrl. (B) MCF7 cells were sequentially transduced with the indicated sgRNAs and with pLX304-empty or pLX304-CUEDC1. The cells were allowed to proliferate for nine days and their growth is represented as percentage of GFP (normalized to T = 0). Data represent mean ± s.d of *n* = 3. ****P* < 0.001. n.s., non-significant. Two-tailed Student's t-test relative to sgCtrl-empty. (C) Western blot analysis of MCF7 cells expressing pLX304-empty and pLX304-CUEDC1. The V5 tag was used to detect expression of CUEDC1. HSP90 was used as a loading control.

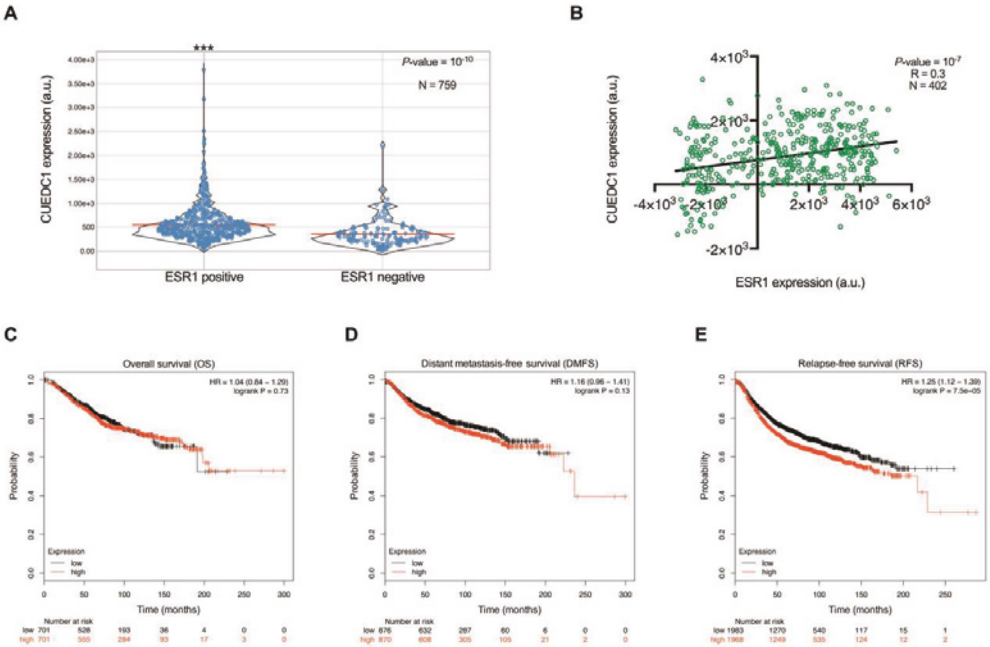


Figure 6 *CUEDC1* expression is significantly increased in ERα-positive tumors and associated with poor clinical outcome of breast cancer patients. (A) The expression of *CUEDC1* was analyzed in breast tumor samples (n = 759; data available on TCGA), which were grouped according to *ESR1* status (positive or negative). ****P* < 0.001 (B) The expression of *CUEDC1* and ESR1 in breast tumor samples (n = 402; data available on TCGA) was used to calculate correlation coefficients (R). R = 0.3; *P*-value = 10^{-7}. (C-E) The prognostic value of *CUEDC1* expression for overall survival (C), distant metastasis-free survival (D) and relapse-free survival (E) of breast cancer patients was analyzed using the Kaplan-Meier plotter tool[34]. The survival curve, hazard ratio (HR) with 95% confidence intervals and logrank *P*-values are displayed.

Methods

Cell lines and chemical reagents. MCF7, MDA-MB-231 and HEK293T cells were cultured in DMEM (Gibco), supplemented with 10% FCS (Hyclone), and 1% penicillin/streptomycin (Gibco). For the estrogen-stimulation experiments, MCF7 cells were cultured in DMEM phenol red–free medium (Gibco) supplemented with 5% charcoal stripped serum (Gibco) for 72h prior to E2 treatment. E2 (17β-estradiol) was purchased from Sigma. All cell lines were obtained from the American Type Culture Collection, and they were tested for mycoplasma.

Cloning of sgRNAs. Custom sgRNAs were designed using CRISPR Design tool (http://crispr.mit.edu/) and cloned into lentiCRISPRv2 (gift from Feng Zhang (Addgene plasmid #52961) according to the protocol described by the Zhang lab[49]. The oligos were purchased from IDT and their sequences are listed on Table 1.

Transduction of human cell lines. To produce lentivirus, 3 x 10^6 HEK293T cells were seeded in 100 mm dishes one day prior to transfection. For each dish, we mixed 10 µg of the target lentiviral construct, 3.5 µg of pVSV-G, 5 µg of pMDL RRE and 2.5 µg of pRSV-REV in a total volume of 450 µl, added 50 µl of CaCl2 and incubated 5 min at RT. Plasmid DNA was precipitated by adding 500 µl 2x HBS to the solution while vortexing at full speed. Lentivirus-containing supernatants were collected 48h post-transfection, filtered through a 0.45 µm membrane (Milipore Steriflip HV/PVDF) and stored at −80 °C. All cell types and lentivirus batches tested were titrated in order to achieve a multiplicity of infection (MOI) of ~0.3. Cell lines were infected with lentivirus supernatants supplemented with 8 µg/ml polybrene (Sigma). At 24 h post-infection, medium was replaced and cells were selected with 2 µg/ml puromycin (Gibco) until there were no cells surviving on the negative control plate (non-transduced cells).

DNA-seq. Cell pellets were collected and gDNA was isolated with DNeasy Blood and Tissue kit (Qiagen). For each sample, we used 500 ng of gDNA as input for the first PCR. We used 5 µl of PCR product as input for the second PCR. Amplification was carried out with 18 cycles for both first and second PCR. After the second PCR, amplicons were purified using Agencourt AMPure XP beads (Beckman Coulter), quantified in a Bioanalyzer 2100 (Agilent), and sequenced using a HiSeq 2500 (Illumina).

RNA isolation, reverse transcription and qPCR. RNA was harvested using TriSure (Bioline) reagent for cell lysis and Rneasy mini kit (Qiagen) to isolate total RNA according to the manufacturer's protocols. cDNA was produced from RNA using Superscript III (Invitrogen). qPCR experiments were performed in a Lightcycler 480 (Roche) using SensiFAST SYBR (Bioline) according to the manufacturer's protocol. The primers used for this experiment are listed on Table 2.

Western blot. Whole-cell lysates were prepared as previously described[25]. Membranes were immunoblotted with the following antibodies: V5 (ab27671, Abcam); HSP90 (H-114, Santa Cruz Biotechnology). Protein bands were visualized using corresponding secondary antibodies (Dako) and ECL reagent (GE Healthcare).

Cell proliferation assay. MCF7 cells were transduced with lentiCRISPRv2 containing sgRNAs of interest (listed on Table 1). Separately, we generated MCF7 cells stably expressing GFP using pLX304-GFP (gift from David Root; Addgene plasmid #25890). GFP-expressing cells were mixed with cells containing individual lentiviral constructs in a 1:3 ratio. The percentage of GFP-expressing cells was assessed by flow cytometry at the beginning of the experiment (T = 0) and at subsequent time-points. We recorded at a minimum of 10,000 events for each condition, and the data were analyzed using FlowJo software.

Luciferase reporter assays. A DNA fragment (~1 kb) containing CUTE (hg19 assembly-chr17:55,976,833-55,977,839) was amplified by PCR from gDNA of MCF7 cells and cloned into pGL3-promoter using KpnI/NheI. The constructs were transfected into MCF-7 cells and treated with 10^{-8} M E2 or vehicle (DMSO) for 24 h. Reporter activity was measured 40 h after transfection using Dual-Luciferase system (Promega) according to the manufacturer's instructions.

RNA-seq. RNA-seq samples were processed with TruSeq RNA library prep kit v2 (Illumina) and sequenced in a HiSeq 2500 (Illumina). Sequenced reads were aligned to the human genome (hg19) using TopHat2[50] and gene expression counts were calculated using HTseq34 based on Ensembl's human gene annotations (v69)[51]. Gene expression levels were normalized by quantiles.

ChIP-seq experiments and data analysis. ChIP-seq of ERα, FOXA1 and H3K27ac in WT and mutant MCF7 cells were performed as described before[52] using the following antibodies: ERα (SC-543; Santa Cruz Biotechnology); FOXA1 (SC-6554; Santa Cruz Biotechnology); H3K27ac (39133; Active Motif). Sequencing reads were aligned to the human genome (hg19) using bwa v 0.7.5 with default parameters. Peaks in control MCF7 cells were called with MACS36 (default parameters) and DFilter37 (parameters: −bs = 50 −ks = 30 −refine −nonzero) algorithms. In-house MCF7 mixed input was used for peak calling of MCF7 cell line. Intersect of the two peak calling algorithms was used for further analysis. ChIP-seq data sets of ERα-Vehicle and ERα-E2 in MCF7 cells was obtained from GEO. The raw files were aligned to hg19 using Bowtie[53]. Unique reads were converted into bigWig files using BEDTools[54] for visualization in the UCSC Genome Browser.

ChIA-PET data. ChIA-PET data (publicly available on the Washington University Epigenome browser) of ERα and RNAPII in MCF7 cells was reanalyzed to identify long-range chromatin interactions at the CUTE locus.

DNase-seq data. DNase-seq data sets from the ENCODE project were reanalyzed to identify open chromatin regions at the CUTE locus. Available tracks were uploaded to the UCSC genome browser for visualization.

GRO-seq experiments and data analysis. GRO-seq experiments in MCF7 and MDA-MB-231 cells were performed as described before[25]. We also used publicly available GRO-seq data of MCF7 cells treated with E2 that was obtained from[14]. Data points were downloaded from the UCSC genome browser using table browser. Transcription levels were quantified by calculating of the sum of data points per gene and normalizing to their length.

Analysis of gene expression and patient survival in breast tumor samples. The Regulome Explorer tool (explorer.cancerregulome.org) was used analyze the expression of *CUEDC1* in breast tumor samples from TCGA (ID: BRCA). The prognostic value of *CUEDC1* expression was analyzed in publicly available breast cancer datasets using the Kaplan-Meier plotter tool (probe 219468_s_at)[34].

Table 1. Sequences of sgRNAs

Name	Oligo 1 (forward)	Oligo 2 (reverse)
sg1830	caccgtttacagcattggtaaggtc	aaacgaccttaccaatgctgtaaac
sgCUEDC1 exon#1	caccgaccacatgcacgtgttcgac	aaacgtcgaacacgtgcatgtggtc
sgCUEDC1 exon#2	caccgtaggctggcggggagtacac	aaacgtgtactccccgccagcctac
sgCUEDC1 intron#1	caccgtaacaagagtttgaactgcg	aaaccgcagttcaaactcttgttac
sgCUEDC1 intron#2	caccgtaaggcttgaggtcaacgat	aaacatcgttgacctcaagccttac

Table 2. Primers used for qPCR experiments

CUEDC1 Forward	aaggaactgcaacggaacc
CUEDC1 Reverse	ggattcgtatttcaatcgatctct
CUTE Forward	acaccagcttcctggttcc
CUTE Reverse	ctgaggtccttccctgcac
TBP Forward	ggagagttctgggattgtac
TBP Reverse	cttatcctcatgattaccgcag

References

1 Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* **136**, E359-386, doi:10.1002/ijc.29210 (2015).

2 Hayashi, S. I. *et al.* The expression and function of estrogen receptor alpha and beta in human breast cancer and its clinical application. *Endocr Relat Cancer* **10**, 193-202 (2003).

3 Sanchez, R., Nguyen, D., Rocha, W., White, J. H. & Mader, S. Diversity in the mechanisms of gene regulation by estrogen receptors. *Bioessays* **24**, 244-254, doi:10.1002/bies.10066 (2002).

4 Arpino, G. *et al.* Molecular mechanism and clinical implications of endocrine therapy resistance in breast cancer. *Oncology* **77 Suppl 1**, 23-37, doi:10.1159/000258493 (2009).

5 Fabian, C. J. The what, why and how of aromatase inhibitors: hormonal agents for treatment and prevention of breast cancer. *Int J Clin Pract* **61**, 2051-2063, doi:10.1111/j.1742-1241.2007.01587.x (2007).

6 Early Breast Cancer Trialists' Collaborative, G. *et al.* Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *Lancet* **378**, 771-784, doi:10.1016/S0140-6736(11)60993-8 (2011).

7 Prall, O. W., Rogan, E. M., Musgrove, E. A., Watts, C. K. & Sutherland, R. L. c-Myc or cyclin D1 mimics estrogen effects on cyclin E-Cdk2 activation and cell cycle reentry. *Mol Cell Biol* **18**, 4499-4508 (1998).

8 Carroll, J. S. *et al.* Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* **38**, 1289-1297, doi:10.1038/ng1901 (2006).

9 Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**, 958-970, doi:10.1016/j.cell.2008.01.018 (2008).

10 Levine, M. Transcriptional enhancers in animal development and evolution. *Curr Biol* **20**, R754-763, doi:10.1016/j.cub.2010.06.070 (2010).

11 Sur, I. & Taipale, J. The role of enhancers in cancer. *Nat Rev Cancer* **16**, 483-493, doi:10.1038/nrc.2016.62 (2016).

12 Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**, 93-108, doi:10.1038/nrg.2015.17 (2016).

13 Hah, N., Murakami, S., Nagari, A., Danko, C. G. & Kraus, W. L. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res* **23**, 1210-1223, doi:10.1101/gr.152306.112 (2013).

14 Li, W. *et al.* Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516-520, doi:10.1038/nature12210 (2013).

15 Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187, doi:10.1038/nature09033 (2010).

16 Koch, F. *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat Struct Mol Biol* **18**, 956-963, doi:10.1038/nsmb.2085 (2011).

17 Hah, N. *et al.* A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145**, 622-634, doi:10.1016/j.cell.2011.03.042 (2011).

18 Melo, C. A. *et al.* eRNAs are required for p53-dependent enhancer activity and gene transcription. *Mol Cell* **49**, 524-535, doi:10.1016/j.molcel.2012.11.021 (2013).

19 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821, doi:10.1126/science.1225829 (2012).

20 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823, doi:10.1126/science.1231143 (2013).

21 Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823-826, doi:10.1126/science.1232033 (2013).

22 Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87, doi:10.1126/science.1247005 (2014).

23 Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84, doi:10.1126/science.1246981 (2014).

24 Diao, Y. *et al.* A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res* **26**, 397-405, doi:10.1101/gr.197152.115 (2016).

25 Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol* **34**, 192-198, doi:10.1038/nbt.3450 (2016).

26 Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat Biotechnol* **34**, 167-174, doi:10.1038/nbt.3468 (2016).

27 Sanjana, N. E. *et al.* High-resolution interrogation of functional elements in the noncoding genome. *Science* **353**, 1545-1549, doi:10.1126/science.aaf7613 (2016).

28 Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936, doi:10.1073/pnas.1016071107 (2010).

29 Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318, doi:10.1038/ng1966 (2007).

30 Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858, doi:10.1038/nature07730 (2009).

31 Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299-308 (1981).

32 Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D. & Carroll, J. S. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat Genet* **43**, 27-33, doi:10.1038/ng.730 (2011).

33 Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat Rev Genet* **14**, 288-295, doi:10.1038/nrg3458 (2013).

34 Gyorffy, B. *et al.* An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res Treat* **123**, 725-731, doi:10.1007/s10549-009-0674-9 (2010).

35 Hah, N. & Kraus, W. L. Hormone-regulated transcriptomes: lessons learned from estrogen signaling pathways in breast cancer cells. *Mol Cell Endocrinol* **382**, 652-664, doi:10.1016/j.mce.2013.06.021 (2014).

36 Lopes, R., Korkmaz, G. & Agami, R. Applying CRISPR-Cas9 tools to identify and characterize transcriptional enhancers. *Nat Rev Mol Cell Biol* **17**, 597-604, doi:10.1038/nrm.2016.79 (2016).

37 Ross-Innes, C. S. *et al.* Cooperative interaction between retinoic acid receptor-alpha and estrogen receptor in breast cancer. *Genes Dev* **24**, 171-182, doi:10.1101/gad.552910 (2010).

38 Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947, doi:10.1016/j.cell.2013.09.053 (2013).

39 Loven, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320-334, doi:10.1016/j.cell.2013.03.036 (2013).

40 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).

41 Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13-23, doi:10.1016/j.cell.2017.02.007 (2017).

42 Dukler, N., Gulko, B., Huang, Y. F. & Siepel, A. Is a super-enhancer greater than the sum of its parts? *Nat Genet* **49**, 2-3, doi:10.1038/ng.3759 (2016).

43 Pott, S. & Lieb, J. D. What are super-enhancers? *Nat Genet* **47**, 8-12, doi:10.1038/ng.3167 (2015).

44 Shih, S. C. *et al.* A ubiquitin-binding motif required for intramolecular monoubiquitylation, the CUE domain. *EMBO J* **22**, 1273-1281, doi:10.1093/emboj/cdg140 (2003).

45 Bagola, K. *et al.* Ubiquitin binding by a CUE domain regulates ubiquitin chain formation by ERAD E3 ligases. *Mol Cell* **50**, 528-539, doi:10.1016/j.molcel.2013.04.005 (2013).

46 Zhang, X. *et al.* An Interaction Landscape of Ubiquitin Signaling. *Mol Cell* **65**, 941-955 e948, doi:10.1016/j.molcel.2017.01.004 (2017).

47 Komander, D. & Rape, M. The ubiquitin code. *Annu Rev Biochem* **81**, 203-229, doi:10.1146/annurev-biochem-060310-170328 (2012).

48 Biewenga, P. *et al.* Gene expression in early stage cervical cancer. *Gynecol Oncol* **108**, 520-526, doi:10.1016/j.ygyno.2007.11.024 (2008).

49 Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods* **11**, 783-784, doi:10.1038/nmeth.3047 (2014).

50 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36, doi:10.1186/gb-2013-14-4-r36 (2013).

51 Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res* **43**, D662-669, doi:10.1093/nar/gku1010 (2015).

52 Zwart, W. *et al.* Oestrogen receptor-co-factor-chromatin specificity in the transcriptional regulation of breast cancer. *EMBO J* **30**, 4764-4776, doi:10.1038/emboj.2011.368 (2011).

53 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).

54 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

Chapter 6

General discussion

Genome-wide mapping of regulatory elements by GRO-seq

At present date, putative regulatory elements are commonly identified through a variety of biochemical, genetic and evolutionary methods[1]. High-throughput biochemical techniques include ChIP-seq[2,3], DNase-seq[4] and MPRA[5,6]. However, they have important limitations that restrict their application in identifying regulatory elements. For example, ChIP-seq requires high-affinity antibodies for TFs or histone modifications, and must be performed separately for each target. On the other hand, high-throughput reporter assays like STARR-seq are able to identify elements that are inactive *in situ* because they are independent of native chromatin context[7]. In recent years, it became clear that active regulatory elements are transcribed by RNAPII[8-10]. We describe a detailed protocol of GRO-seq in **chapter 3**, which is a variant of nuclear run-on assays coupled with NGS. This technique has high sensitivity and allows detection of transcribed regions that are not readily identified by measuring steady-state RNA levels[8]. GRO-seq can be used for identifying unannotated transcripts that are responsive to signaling cues, which is a significant advantage for detecting novel regulatory RNAs[11-18]. In our work, we used GRO-seq to identify genome-wide transcriptional changes mediated by p53[19]. As expected, we observed activation of several canonical target genes of p53 (e.g. *CDKN1A*), as well as known eRNAs regulated by p53[20]. Bioinformatic analysis identified 50,502 putative enhancers, of which 6,270 were differentially transcribed upon p53 activation. The intersection of GRO-seq and ChIP-seq data resulted in the identification of enhancers that are regulated by p53. Surprisingly, we also found a large group of p53-activated enhancers that were not associated with binding of p53. The regulatory mechanisms of many of these regions remain elusive, and further studies are required to elucidate whether they have a biological function in the p53 pathway.

GRO-cap is a variant of the original GRO-seq method that includes an enrichment step for 5′-capped RNAs, which further improves the sensitivity and specificity to detect nascent RNAs[21]. Nonetheless, both GRO-seq and GRO-cap rely on the assumption that active regulatory elements are transcribed - which is a major limitation of these techniques[18]. According to GRO-seq data the number of active regulatory elements in a given cell-type is in the order of tens of thousands[11-13,15,16,19]. This estimation is in agreement with ChIP-seq and DNase-seq data[22-25], although there is great variability between studies likely due to cell-type specificity. Of note, active regulatory elements identified by GRO-seq are also enriched in H3K27Ac, binding of TFs and eQTLs[18], suggesting they are actively involved in transcriptional regulation. The proportion of the human genome assigned to candidate biological functions varies markedly among the different approaches, with estimates from biochemical studies being considerably larger than those of genetic and evolutionary analysis[1,26,27]. These contradictory results highlight the importance of integrating different approaches and developing new high-throughput technologies to characterize the functional DNA elements of the human genome.

The role of eRNAs in gene regulation

In **chapter 4**, we used GRO-seq to measure eRNA expression and infer the location of enhancers in a genome-wide scale[28]. This approach is supported by evidence from different studies showing that eRNA expression is a robust and independent indicator of enhancer activity[10,13,15,29-32]. For example, transcribed enhancers exhibit higher levels of chromatin accessibility, binding of transcriptional coactivators and active histone marks[15,29,31,33] compared to non-transcribed enhancers. Additionally, enhancers that produce eRNAs are more likely to display transcriptional activity in reporter assays than non-transcribed enhancers[10]. It was also shown that transcriptional regulatory elements can be predicted solely through eRNA expression without using chromatin marks[18,21]. These results illustrate the predictive power of eRNA expression for annotating enhancers, and also support a role for these transcripts in gene regulation.

Data from different studies suggest that there are ~40,000-65,000 eRNAs[10,34], which constitutes a large faction of transcription events in human cells[35]. Still, it is unclear what is the biological significance of enhancer transcription and how eRNAs contribute to the activation of gene expression. To date, several mechanisms were proposed to answer these questions, including regulation of chromatin accessibility[36], TF binding[37], RNAPII loading[36,38] and pause release[39]. There is also evidence supporting a role for eRNAs in the initiation or maintenance of enhancer-promoter looping[13,40]. However, several studies found no effect on enhancer-promoter interactions upon knockdown of eRNAs[20,39] or chemical inhibition of RNAPII elongation[15]. Moreover, the contribution of eRNAs to the recruitment of mediator and cohesin remains elusive. These conflicting results might reflect different mechanisms at specific enhancers and target genes, and prompt further studies to clarify the role of eRNAs in DNA looping. This question can be addressed by performing tethering experiments using dCas9-eRNA fusions and assessing enhancer-promoter interactions. Future studies should also aim at identifying the protein partners of functional eRNAs[41,42] in order to obtain detailed biochemical insights into their mechanisms of action.

Another open question in the field is whether eRNAs activate transcription in *cis* or in *trans*. The fact that eRNAs are (1) usually lowly expressed, (2) predominantly act on neighboring genes and (3) mainly localize to the nucleus supports the *cis* hypothesis. Li *et al*. performed ChIRP-seq of an estrogen-induced eRNA (FOXC1-eRNA), but failed to identify any target genes in *trans*[13]. However, there are several examples of lncRNAs (e.g. *Jpx*[43] and *NeST*[44]) that activate their targets in *trans*, suggesting that this might also be the case for eRNAs. Indeed, two studies found that depletion of eRNAs (KLK3-eRNA and DRR-eRNA) affected the expression of multiple genes, some of which were located in different chromosomes[36,45]. Both KLK3-eRNA and DRR-eRNA are polyadenylated whereas FOXC1-eRNA is not, leading to the hypothesis that eRNAs with relatively high stability can function in *trans*. This is supported by knockdown experiments of enhancer-like lncRNAs, which affected the expression of hundreds of genes[46]. It should be stressed that none of the eRNA functions mentioned above has been proved to be independent of the enhancer sequence. Therefore, additional evidence from imaging experiments (e.g. single-molecule RNA labeling), genetic manipulation (e.g. RNAi and CRISPR-Cas9) and functional assays (e.g., tethering eRNAs to target loci) is required to confirm that these transcripts act in *trans*.

What are super-enhancers?

Genome-wide profiling of H3K27Ac and H3K4me1 identified a putative new class of regulatory elements termed “super-enhancers”. These regions are bound by the Mediator complex, span ~5-50 kb in length and are flanked by CTCF-binding sites[47,48]. According to this definition, there are more than 200 super-enhancers in the human genome, and most of them seem to be associated with critical developmental genes[47]. Since the term was initially proposed in 2013, hundreds of studies were published about the function of super-enhancers in different cell-types and biological settings[49]. Yet, it is not clear whether they constitute a novel paradigm in gene regulation[50]. It is possible that super-enhancers have a similar function to the previously identified LCR and GCR, which coordinate the expression of linked genes in specific cell-types. In fact, the LCRs of α*-globin* and β*-globin* are strikingly similar to super-enhancers: they contain several enhancers, each bound by multiple TFs[51,52], and fit the bioinformatics criteria for super-enhancers[53]. Recent studies addressed the question of how the constituent elements of a super-enhancer contribute to activate the expression of a target gene. Hay *et al*. deleted individual constituents within the α*-globin* super-enhancer and found that they act independently and in an additive manner to regulate gene expression[53]. Shin *et al*. used a similar approach to characterize the *WAP* super-enhancer and found that its individual constituent enhancers have partially redundant roles[54]. Mathematical modeling of these two datasets[53,54] indicates that individual enhancers contribute to the activity of a super-enhancer in a linear function[55]. Altogether, these studies found no evidence of novel functional properties of super-enhancer regions: each element seems to contribute to gene expression as an individual enhancer in an additive rather than synergistic manner.

According to the super-enhancer model, these elements confer stronger activation of gene expression compared to “typical” enhancers[47,48]. Moorthy *et al*. performed a systematic comparison between enhancers and super-enhancers by deleting tens of loci in mouse ESCs. They found that deletion of super-enhancers resulted in highly variable effects on the expression of target genes (decreased levels ranging from 12% to 92%). Moreover, a substantial number of highly transcribed genes in ES cells are not associated with a super-enhancer[56]. In **chapter 4** and **5**, we studied the contribution of individual enhancers to gene expression by generating mutations in their genetic sequence. In all cases, we observed a strong reduction in the expression of their targets (>50%), some of which are known tumor-suppressor genes and oncogenes (e.g. *CDKN1A* and *CCND1*). Despite not fitting the criteria of super-enhancers, these enhancers mediate a strong activation of gene expression and are required for critical phenotypes in human cells such as senescence and G1/S transition. Based on available evidence, it seems that enhancers and super-enhancers have an equivalent role in the regulation of single or multiple genes[56]. Moreover, the super-enhancer model likely ignores a large number of functional enhancers in human cells, and exaggerates the importance of clustered enhancers compared to isolated enhancers.

Identification of disease-causal mutations in the non-coding genome

In **chapter 4**, we studied the function of a distal enhancer of *CDKN1A* (deCDKN1A) by mutating it with CRISPR-Cas9 system. In reporter assays, the transcriptional activity of deCDKN1A is completely abolished by single-nucleotide deletions within the p53-binding motif. This result goes against the well-established notion that enhancer sequences redundant and robust to small mutations[57], and suggests that single-nucleotide alterations can markedly influence gene expression by changing the activity of regulatory elements. This hypothesis is supported by data of GWAS, which identified a large number of SNPs that are associated with cancer and other common diseases[58]. However, the functional outcome of mutations in enhancer elements can differ substantially from that of mutations in protein-coding genes. Mutations in enhancers are largely constrained to effects in *cis*, whereas mutations in protein-coding sequences can alter several aspects of gene regulation - mRNA processing, stability and translation, or even protein structure and activity[59]. In addition, the impact of genetic alterations in regulatory elements is predicted to have less detrimental effects than those in protein-coding genes since their activity is often restricted in space and time[60]. Nonetheless, both germline variants and somatic mutations in enhancers have been causally linked with a range of human diseases[58,61]. It is worth noting that the impact of germline and somatic alterations in gene expression can differ substantially. Usually, germline variants modify the binding affinity of TFs and change mildly the activity of enhancers. This can lead to increased risk of cancer development, without affecting the fitness of the entire organism. In contrast, somatic mutations do not have to be carried through the germline and, therefore, can have a much greater impact on gene expression. This is illustrated by the identification of mutations in enhancers that activate proto-oncogenes and drive cancer development[62].

The majority of genetic alterations in enhancers was identified by targeted DNA sequencing after their function has been characterized; or otherwise, by the identification of large deletions or rearrangements that were subsequently shown to involve enhancer elements. These approaches are relatively inefficient compared with exome sequencing, which is very successful in detecting recurrent mutations in protein-coding genes. The large amount of evidence that is required to assert causality to a non-coding variant remains a great challenge in biomedical research. In recent years, the application of whole-genome sequencing led to the discovery of recurrent mutations in regulatory elements in different types of tumors[63], and this tendency is likely to increase as the costs of NGS technologies continue to fall[64]. However, as in classic Mendelian diseases involving protein-coding, it is possible to implicate a non-coding variant by combining multiple lines of evidence, including human genetics (e.g. fine mapping, *trans*-ethnic studies) and functional studies (e.g. reporter assays, genome editing). Ultimately, the goal is not solely to determine which variants are causal, but also to understand the genetic pathways they regulate and harness this knowledge to develop more efficient therapies.

A clash of titans: RNAi vs CRISPR-Cas9

The success of RNAi technologies is illustrated by more than ninety thousand studies published since the discovery of gene silencing in *C. elegans* [65,66]. RNAi reagents can be used to target both coding and non-coding RNAs and produce knockdown of gene expression. However, they frequently yield false-negative results due to inefficient knockdown of the target RNA. In addition, the high prevalence of off-target effects, where additional genes are unintentionally perturbed, leads to false-positive results. Such issues are the most probable cause of poor reproducibility between RNAi screens[67], and considerable effort has been made to improve both methods and reagents. For example, current shRNA libraries contain many independent constructs (up to 25) targeting each gene[68]. In this way, it is possible to overcome limitations of individual reagents to identify high-confidence candidates by pooling the results of each of them. Nevertheless, working with ultra-complex libraries is a very laborious process and requires performing extensive follow-up analyses to identify the most robust candidates. In **chapter 4** and **5**, we relied extensively on CRISPR-Cas9 to study both protein-coding genes and non-coding regulatory elements. In contrast to RNAi, CRISPR-Cas9 reagents often generate loss-of-function mutations in the target cells[69-72].

Moreover, they circumvent a major limitation of RNAi libraries due to the capacity of Cas9 to induce stable and heritable mutations on target DNA sequences. The fact that Cas9 can be targeted to virtually any genomic sequence makes CRISPR systems very suitable to the function of regulatory elements[73]. We and others have used CRISPR-Cas9 to disrupt the genetic sequence of enhancers and screen for various phenotypes (e.g. cell proliferation[28] and drug resistance[74]). Alternatively, CRISPR systems based on fusions of dCas9 with activator (e.g. p300[75,76]) or repressor (e.g. KRAB[77,78]) molecules can be used to perform gain-of-function or loss-of-function studies, respectively. The ability to perform gain-of-function experiments, such as activation of enhancers by dCas9-p300[75,76], is a major advantage of CRISPR compared to RNAi systems. Nevertheless, both siRNA and shRNA are invaluable tools to target transcripts derived from regulatory regions, such as lncRNAs[46] and eRNAs[13,20]. To date, there are no large-scale RNAi libraries available to perform genetic screens of lncRNAs or eRNAs. The systematic manipulation of regulatory elements by CRISPR and RNAi tools can elucidate the functions and mechanisms of non-coding DNA sequences and their transcripts.

Recent studies have systematically compared the performance of shRNA and sgRNA libraries in genetic screens. One of them concluded that CRISPR-Cas9 reagents have higher consistency and lower off-target effects compared to RNAi [79]. The other concluded that both technologies perform similarly, although it was noted that CRISPR-Cas9 identified many more essential genes using a much smaller library[80]. It is clear that both RNAi and CRISPR-Cas9 technologies have sensitivity and specificity issues, but the combination of both screening methods can better discriminate positive from negative control genes. Indeed, parallel genome-wide knockdown and knockout screens were successfully used to probe the mechanism of action of anti-viral drugs[81]. These results demonstrate that the combined application of both technologies is more powerful than either alone because each method identifies only a subset of relevant genes[81]. This suggests that combining RNAi and CRISPR-Cas9 can be a valuable method for screening complex phenotypes, such as the identification of therapeutic targets or developmental processes.

A major limitation of current CRISPR-Cas9 systems is the requirement of a specific PAM, immediately downstream of the target site, for efficient gene editing[82,83]. The most widely used system is the type II CRISPR from *S. pyogenes* which requires the presence of a PAM in the form of NGG. As a result, it is often difficult to target DSBs with precision for specific genome editing applications like HDR and non-coding genetic screens. The work described on **chapter 4** clearly illustrates this problem, as we were able to target only a fraction of p53- and ERα-bound enhancers (90% and 60%, respectively)[28]. The number of TF binding sites that are targetable by Cas9 can be even lower for DNA motifs poor in C/G nucleobases (e.g. GATA3 and FOXA1). Two main strategies have been used to broaden the targeting range of Cas9: engineering of CRISPR systems from bacterial species that have different PAM requirements than *S. pyogenes*[84-86]; and altering the PAM requirements of Cas9 proteins by directed molecular evolution[87,88]. Recently, Cpf1 was identified as an additional type II CRISPR endonuclease that requires a TTN PAM[89,90], providing an attractive alternative to target T-rich sequences. Moreover, Hirano *et al*. created a variant of *F. novicida* Cas9 that recognizes a shorter and less stringent PAM (YG)[91]. The identification or creation of novel nucleases that require smaller PAMs will dramatically increase the applications of CRISPR systems and broaden the scope of genome-wide genetic screens.

Towards a comprehensive identification of functional regulatory elements in the human genome

The identification of functional enhancers and their target genes is a major challenge in the post-Human Genome Project era. In **chapter 4**, we showed that it is possible to identify enhancers and characterize their function in an unprecedented scale. We started by identifying putative enhancers through histone marks, eRNA transcription and TF binding. This yielded a confident set of candidate enhancers that are active in a specific biological setting. Second, we built CRISPR-Cas9 libraries to screen putative enhancers in a high-throughput fashion. This allowed us testing multiple enhancers and their contribution to a specific phenotype. Of note, this approach proved to be successful in identifying functional enhancers in both enrichment and dropout genetic screens. Finally, we used differential gene expression profiling by RNA-seq to identify the targets of each enhancer. The usefulness of this method is highlighted in **chapter 5**, where we characterized the regulation of *CUEDC1* expression by ERα. In MCF7 cells, ERα binds to more than 14,000 loci[92] and regulates the expression of ~3,000 protein-coding genes[15]. These genome-wide studies revealed that ERα regulates gene expression by binding mostly to distal enhancers.

However, they cannot elucidate the function of these enhancers nor pinpoint the primary targets of estrogen. Therefore, it is not surprising that only a few number of genes was shown to be required for ERα-mediated cell proliferation (e.g. *CCND1* and *MYC*)[93]. We identified CUTE in a CRISPR-Cas9 screen for enhancers that are essential for the proliferation of ERα-positive cells[28]. Following the steps mentioned above, we found that ERα activates the expression of *CUEDC1* through CUTE. *CUEDC1* is a fairly unknown gene and not much is known about its biological functions so far. Our work revealed that *CUEDC1* is required for cell proliferation mediated by ERα and is a potential biomarker of breast cancer. These findings highlight the power of genetic manipulation by CRISPR-Cas9 to determine the function of regulatory elements and genes that were previously uncharacterized.

It is worth noting that our work is a proof-of-concept that has intrinsic limitations. For example, the first step of our approach relied on the assumption that enhancers contain specific genetic and epigenetic marks, such as TF binding motifs and histone modifications. Therefore, we likely missed candidate enhancers simply because they did not contain these features. This limitation is underlined by the findings of Rajagopal *et al.*, which identified active regulatory elements that are not marked by known biochemical features[94]. Additionally, the design of sgRNAs is constrained by the availability of PAMs in the target loci. The magnitude of non-coding CRISPR-Cas9 screens performed to date is far from comprehensive, as they were confined to regions of ~2 kb to ~1 Mb and tested the function of a few thousand candidates[73]. The identification of *all* regulatory elements in the human genome is a monumental task, which will probably require technological improvements (e.g. tools to efficiently generate large deletions), increased speed of screening (e.g. using robotics) and, most importantly, collaborative work between different research institutes (e.g. a large research consortium like the Human Genome Project). The exact number of functional enhancers in the human genome is debatable, but it is predicted that the human genome contains roughly one million enhancers[24,25]. A back-of-the-envelope calculation indicates that there are ~20.000 enhancers in the smallest autosome, whereas the biggest contains ~80.000 (assuming that chromosome 1 and 22 correspond to ~8% and ~2% of the total genomic DNA, respectively; and that enhancers are evenly distributed throughout the genome). Therefore, it is completely feasible to perform genetic screens for all putative enhancers on a chromosome-wide scale at present date - as a comparison the human genome has ~22.000 protein-coding genes[95]. This type of experiments has the potential to identify enhancers in a comprehensive manner, and also reveal the elusive mechanisms that underlie their biological functions. I think that this information is invaluable *per se*, and can contribute decisively to the ultimate goal of mankind - defeating human disease.

References

1 Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* **111**, 6131-6138, doi:10.1073/pnas.1318948111 (2014).

2 Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858, doi:10.1038/nature07730 (2009).

3 Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49, doi:10.1038/nature09906 (2011).

4 Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-322, doi:10.1016/j.cell.2007.12.014 (2008).

5 Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**, 271-277, doi:10.1038/nbt.2137 (2012).

6 Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**, 265-270, doi:10.1038/nbt.2136 (2012).

7 Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074-1077, doi:10.1126/science.1232542 (2013).

8 Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845-1848, doi:10.1126/science.1162228 (2008).

9 Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849-1851, doi:10.1126/science.1162253 (2008).

10 Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461, doi:10.1038/nature12787 (2014).

11 Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474**, 390-394, doi:10.1038/nature10006 (2011).

12 Hah, N. *et al.* A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145**, 622-634, doi:10.1016/j.cell.2011.03.042 (2011).

13 Li, W. *et al.* Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516-520, doi:10.1038/nature12210 (2013).

14 Kaikkonen, M. U. *et al.* Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Mol Cell* **51**, 310-325, doi:10.1016/j.molcel.2013.07.010 (2013).

15 Hah, N., Murakami, S., Nagari, A., Danko, C. G. & Kraus, W. L. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res* **23**, 1210-1223, doi:10.1101/gr.152306.112 (2013).

16 Allen, M. A. *et al.* Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. *Elife* **3**, e02200, doi:10.7554/eLife.02200 (2014).

17 Chae, M., Danko, C. G. & Kraus, W. L. groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. *BMC Bioinformatics* **16**, 222, doi:10.1186/s12859-015-0656-3 (2015).

18 Danko, C. G. *et al.* Identification of active transcriptional regulatory elements from GRO-seq data. *Nat Methods* **12**, 433-438, doi:10.1038/nmeth.3329 (2015).

19 Leveille, N. *et al.* Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA. *Nat Commun* **6**, 6520, doi:10.1038/ncomms7520 (2015).

20 Melo, C. A. *et al.* eRNAs are required for p53-dependent enhancer activity and gene transcription. *Mol Cell* **49**, 524-535, doi:10.1016/j.molcel.2012.11.021 (2013).

21 Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46**, 1311-1320, doi:10.1038/ng.3142 (2014).

22 Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318, doi:10.1038/ng1966 (2007).

23 Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-112, doi:10.1038/nature07829 (2009).

24 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

25 Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82, doi:10.1038/nature11232 (2012).

26 Ponting, C. P. & Hardison, R. C. What fraction of the human genome is functional? *Genome Res* **21**, 1769-1776, doi:10.1101/gr.116814.110 (2011).

27 Rands, C. M., Meader, S., Ponting, C. P. & Lunter, G. 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet* **10**, e1004525, doi:10.1371/journal.pgen.1004525 (2014).

28 Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol* **34**, 192-198, doi:10.1038/nbt.3450 (2016).

29 Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187, doi:10.1038/nature09033 (2010).

30 De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol* **8**, e1000384, doi:10.1371/journal.pbio.1000384 (2010).

31 Melgar, M. F., Collins, F. S. & Sethupathy, P. Discovery of active enhancers through bidirectional expression of short transcripts. *Genome Biol* **12**, R113, doi:10.1186/gb-2011-12-11-r113 (2011).

32 Lam, M. T. *et al.* Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature* **498**, 511-515, doi:10.1038/nature12209 (2013).

33 Zhu, Y. *et al.* Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Res* **41**, 10032-10043, doi:10.1093/nar/gkt826 (2013).

34 Arner, E. *et al.* Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**, 1010-1014, doi:10.1126/science.1259418 (2015).

35 Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108, doi:10.1038/nature11233 (2012).

36 Mousavi, K. *et al.* eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol Cell* **51**, 606-617, doi:10.1016/j.molcel.2013.07.022 (2013).

37 Sigova, A. A. *et al.* Transcription factor trapping by RNA in gene regulatory elements. *Science* **350**, 978-981, doi:10.1126/science.aad3346 (2015).

38 Maruyama, A., Mimura, J. & Itoh, K. Non-coding RNA derived from the region adjacent to the human HO-1 E2 enhancer selectively regulates HO-1 gene induction by modulating Pol II binding. *Nucleic Acids Res* **42**, 13599-13614, doi:10.1093/nar/gku1169 (2014).

39 Schaukowitch, K. *et al.* Enhancer RNA facilitates NELF release from immediate early genes. *Mol Cell* **56**, 29-42, doi:10.1016/j.molcel.2014.08.023 (2014).

40 Lai, F. *et al.* Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* **494**, 497-501, doi:10.1038/nature11884 (2013).

41 Chu, C. *et al.* Systematic discovery of Xist RNA binding proteins. *Cell* **161**, 404-416, doi:10.1016/j.cell.2015.03.025 (2015).

42 McHugh, C. A. *et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**, 232-236, doi:10.1038/nature14443 (2015).

43 Tian, D., Sun, S. & Lee, J. T. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell* **143**, 390-403, doi:10.1016/j.cell.2010.09.049 (2010).

44 Gomez, J. A. *et al.* The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-gamma locus. *Cell* **152**, 743-754, doi:10.1016/j.cell.2013.01.015 (2013).

45 Hsieh, C. L. *et al.* Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. *Proc Natl Acad Sci U S A* **111**, 7319-7324, doi:10.1073/pnas.1324151111 (2014).

46 Orom, U. A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58, doi:10.1016/j.cell.2010.09.001 (2010).

47 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).

48 Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947, doi:10.1016/j.cell.2013.09.053 (2013).

49 Niederriter, A. R., Varshney, A., Parker, S. C. & Martin, D. M. Super Enhancers in Cancers, Complex Disease, and Developmental Disorders. *Genes (Basel)* **6**, 1183-1200, doi:10.3390/genes6041183 (2015).

50 Pott, S. & Lieb, J. D. What are super-enhancers? *Nat Genet* **47**, 8-12, doi:10.1038/ng.3167 (2015).

51 Grosveld, F., van Assendelft, G. B., Greaves, D. R. & Kollias, G. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* **51**, 975-985 (1987).

52 Higgs, D. R. *et al.* A major positive regulatory region located far upstream of the human alpha-globin gene locus. *Genes Dev* **4**, 1588-1601 (1990).

53 Hay, D. *et al.* Genetic dissection of the alpha-globin super-enhancer in vivo. *Nat Genet* **48**, 895-903, doi:10.1038/ng.3605 (2016).

54 Shin, H. Y. *et al.* Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nat Genet* **48**, 904-911, doi:10.1038/ng.3606 (2016).

55 Dukler, N., Gulko, B., Huang, Y. F. & Siepel, A. Is a super-enhancer greater than the sum of its parts? *Nat Genet* **49**, 2-3, doi:10.1038/ng.3759 (2016).

56 Moorthy, S. D. *et al.* Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Res* **27**, 246-258, doi:10.1101/gr.210930.116 (2017).

57 Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7**, 29-59, doi:10.1146/annurev.genom.7.080505.115623 (2006).

58 Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**, 93-108, doi:10.1038/nrg.2015.17 (2016).

59 Sauna, Z. E. & Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet* **12**, 683-691, doi:10.1038/nrg3051 (2011).

60 Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246-1250, doi:10.1126/science.1174148 (2009).

61 Krijger, P. H. & de Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nat Rev Mol Cell Biol* **17**, 771-782, doi:10.1038/nrm.2016.138 (2016).

62 Mansour, M. R. *et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373-1377, doi:10.1126/science.1259037 (2014).

63 Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet* **47**, 818-821, doi:10.1038/ng.3335 (2015).

64 Erlich, Y. A vision for ubiquitous sequencing. *Genome Res* **25**, 1411-1416, doi:10.1101/gr.191692.115 (2015).

65 Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806-811, doi:10.1038/35888 (1998).

66 Timmons, L. & Fire, A. Specific interference by ingested dsRNA. *Nature* **395**, 854, doi:10.1038/27579 (1998).

67 Kaelin, W. G., Jr. Molecular biology. Use and abuse of RNAi to study mammalian gene function. *Science* **337**, 421-422, doi:10.1126/science.1225787 (2012).

68 Kampmann, M. *et al.* Next-generation libraries for robust RNA interference-based genome-wide screens. *Proc Natl Acad Sci U S A* **112**, E3384-3391, doi:10.1073/pnas.1508821112 (2015).

69 Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823-826, doi:10.1126/science.1232033 (2013).

70 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823, doi:10.1126/science.1231143 (2013).

71 Hwang, W. Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat Biotechnol* **31**, 227-229, doi:10.1038/nbt.2501 (2013).

72 Jinek, M. *et al.* RNA-programmed genome editing in human cells. *Elife* **2**, e00471, doi:10.7554/eLife.00471 (2013).

73 Lopes, R., Korkmaz, G. & Agami, R. Applying CRISPR-Cas9 tools to identify and characterize transcriptional enhancers. *Nat Rev Mol Cell Biol* **17**, 597-604, doi:10.1038/nrm.2016.79 (2016).

74 Sanjana, N. E. *et al.* High-resolution interrogation of functional elements in the noncoding genome. *Science* **353**, 1545-1549, doi:10.1126/science.aaf7613 (2016).

75 Hilton, I. B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat Biotechnol* **33**, 510-517, doi:10.1038/nbt.3199 (2015).

76 Klann, T. S. *et al.* CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat Biotechnol*, doi:10.1038/nbt.3853 (2017).

77 Kearns, N. A. *et al.* Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nat Methods* **12**, 401-403, doi:10.1038/nmeth.3325 (2015).

78 Fulco, C. P. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769-773, doi:10.1126/science.aag2445 (2016).

79 Evers, B. *et al.* CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. *Nat Biotechnol* **34**, 631-633, doi:10.1038/nbt.3536 (2016).

80 Morgens, D. W., Deans, R. M., Li, A. & Bassik, M. C. Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes. *Nat Biotechnol* **34**, 634-636, doi:10.1038/nbt.3567 (2016).

81 Deans, R. M. *et al.* Parallel shRNA and CRISPR-Cas9 screens enable antiviral drug target identification. *Nat Chem Biol* **12**, 361-366, doi:10.1038/nchembio.2050 (2016).

82 Sapranauskas, R. *et al.* The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic Acids Res* **39**, 9275-9282, doi:10.1093/nar/gkr606 (2011).

83 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821, doi:10.1126/science.1225829 (2012).

84 Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci U S A* **109**, E2579-2586, doi:10.1073/pnas.1208507109 (2012).

85 Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat Methods* **10**, 1116-1121, doi:10.1038/nmeth.2681 (2013).

86 Hou, Z. *et al.* Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis. *Proc Natl Acad Sci U S A* **110**, 15644-15649, doi:10.1073/pnas.1313587110 (2013).

87 Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481-485, doi:10.1038/nature14592 (2015).

88 Kleinstiver, B. P. *et al.* Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. *Nat Biotechnol* **33**, 1293-1298, doi:10.1038/nbt.3404 (2015).

89 Zetsche, B. *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759-771, doi:10.1016/j.cell.2015.09.038 (2015).

90 Yamano, T. *et al.* Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. *Cell* **165**, 949-962, doi:10.1016/j.cell.2016.04.003 (2016).

91 Hirano, H. *et al.* Structure and Engineering of Francisella novicida Cas9. *Cell* **164**, 950-961, doi:10.1016/j.cell.2016.01.039 (2016).

92 Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D. & Carroll, J. S. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat Genet* **43**, 27-33, doi:10.1038/ng.730 (2011).

93 Prall, O. W., Rogan, E. M., Musgrove, E. A., Watts, C. K. & Sutherland, R. L. c-Myc or cyclin D1 mimics estrogen effects on cyclin E-Cdk2 activation and cell cycle reentry. *Mol Cell Biol* **18**, 4499-4508 (1998).

94 Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat Biotechnol* **34**, 167-174, doi:10.1038/nbt.3468 (2016).

95 Pertea, M. & Salzberg, S. L. Between a chicken and a grape: estimating the number of human genes. *Genome Biol* **11**, 206, doi:10.1186/gb-2010-11-5-206 (2010).

Appendix

Scope of the thesis

The identification of enhancers and their target genes is a major challenge in the post-Human Genome Project era. Evidence from different studies indicates that active enhancers are transcribed by RNAPII - giving rise to so-called enhancer-associated RNAs (eRNAs). These transcripts are generally lowly-abundant and, therefore, it is challenging to detect them by standard RNA-seq. In **chapter 3** we described a protocol of GRO-seq, which is a variant of nuclear run-on assays coupled with NGS. This technique has higher sensitivity compared to RNA-seq, enabling the detection of transcripts that are not readily identified by measuring steady-state RNA levels. In **chapter 4** we asked if the expression of eRNAs can predict the location of enhancers that are responsive to specific TFs in a genome-wide scale? For that purpose we used GRO-seq to measure the expression of eRNAs and combined that information with ChIP-seq data of p53 and ERα. The data obtained by GRO-seq is complementary to other datasets obtained by NGS (e.g MPRA), but it has an added value of predicting the activity of enhancers in their native context. However, these techniques cannot reveal the contribution of enhancers to specific biological processes. In **chapter 4** we interrogated the function of hundreds of enhancers in parallel by performing genetic screens in human cells. Our experimental strategy is based on the disruption of the sequence of enhancers by CRISPR-Cas9, which led to the identification of enhancers that regulate the expression of oncogenes and tumor-suppressor genes in human cells. Additionally, we showed in **chapter 5** that *CUEDC1* – a previously uncharacterized gene - is regulated by ERα through an intronic enhancer and is required for the proliferation of breast cancer cells. In conclusion, we demonstrated that the combination of gene expression analysis and genetic screening allows characterizing enhancers and their target genes in an unprecedented scale. We propose that this type of experiments can elucidate the function and mechanism of action of enhancers, and ultimately contribute to improve the diagnosis and therapies of human pathologies.

Summary

The sequence of DNA is a code that contains *all* the information that is required for life (as we know it). DNA is stored inside the nucleus of cells and its sequence is replicated during cell division to ensure that the genetic information is transmitted to the daughter cells. The information contained in DNA is copied into RNA by a process called transcription. RNA acts as a messenger (mRNA) to carry the information between the nucleus and the cytoplasm, where it is used as a template to produce proteins through a process called translation. Proteins are the main effectors of all biological functions in the cell. However, the information required to make proteins (called "coding DNA sequence") comprises only a small portion (~2%) of the entire human genome sequence. For several decades, it was generally accepted that the remaining 98% of the genome sequence had no biological function and, because of that, it was dubbed "junk DNA". The discovery of non-coding DNA sequences that control the expression of genes challenged this idea, and revealed that there is biological function beyond protein-coding sequences. These non-coding sequences are called "regulatory elements" and they are classified into four classes according to their function: promoters, enhancers, insulators and silencers. Among them, enhancers play a critical role in activating the expression of genes in response to intra- and extra-cellular stimuli – which is essential for the development of complex organisms. Previous studies suggest that the human genome might contain more than one million enhancers – a much higher number compared to the number of protein-coding genes (~22.000). However, not much is known about the biological function of most enhancers since only a handful of them were studied in detail to present date.

In recent years, the use of next-generation sequencing technologies (e.g. RNA-seq) revealed that the vast majority of the human genome is transcribed into non-coding RNA species. Initially, it was thought that these transcripts were "junk RNA" transcribed from "junk DNA". This hypothesis was refuted by thousands of studies reporting that non-coding RNAs regulate a remarkably broad spectrum of cellular processes – including transcription and translation. Moreover, it was shown that the dysregulation of regulatory elements (and the non-coding RNAs transcribed from them) is associated with different human pathologies such as cancer. In this work, we describe a detailed protocol of Global Run-on sequencing (GRO-seq), which is a high-throughput sequencing technique that measures nascent RNA transcription. We applied GRO-seq to detect enhancer-associated RNAs (eRNAs), which are non-coding RNAs transcribed from active enhancers. Our experiments identified thousands of enhancers that are activated by critical transcription factors (e.g. p53 and ERα) and might play a role in cancer development. In order to characterize their function, we used a recently developed technology called CRISPR-Cas9. This system is composed of a protein that can cleave DNA (Cas9) and a nucleic acid sequence that can guide Cas9 to the target site (CRISPR). CRISPR-Cas9 triggered a revolution in biology because it allows editing specific DNA sequences in a very fast and easy way. Importantly, Cas9 can be directed to virtually any sequence of the human genome, thus allowing to study the function of enhancers and other regulatory elements in a comprehensive manner. We pioneered the application of CRISPR-Cas9 to test the function of enhancers by mutating them and examining the resulting phenotype in a high-throughput manner (i.e. genetic screening). Our experiments led to the identification of several enhancers that regulate the expression of critical genes (e.g. *CCND1*) and poorly-characterized genes (e.g. *CUEDC1*) in human cells. In both cases, we showed that these enhancers are absolutely required for the growth of cancer cells, and our findings provide the basis for better diagnosis and therapies of human cancer.

Samenvatting

De sequentie van DNA is een code die *alle* informatie bevat die benodigd is voor leven (zoals we dat kennen). DNA wordt opgeslagen in de kern van cellen en de sequentie wordt gekopieerd tijdens de celdeling om ervoor te zorgen dat de genetische informatie wordt overgedragen aan de dochtercellen. De informatie die DNA bevat wordt gekopieerd in RNA door een proces genaamd transcriptie. RNA fungeert als boodschapper ("messenger" of "mRNA") om de informatie van de kern naar het cytoplasma van de cel te dragen, waar het gebruikt wordt als sjabloon om eiwitten te produceren, via een proces genaamd translatie. Eiwitten zijn de belangrijkste uitvoerders van biologische functies in de cel. Echter, de informatie om eiwitten te maken beslaat slechts een klein deel van het menselijk genoom (~2%). Gedurende meerdere decennia was het algemeen geaccepteerd dat de resterende ~98% van het menselijk genoom geen functie had, en werd daarom bestempeld als "rommel DNA" ("junk DNA"). De ontdekking van niet-coderende DNA sequenties die de expressie van genen kunnen reguleren zorgde dat dit idee werd betwist en onthulde dat DNA sequenties die niet voor eiwit coderen ook functies kunnen hebben. Deze niet-coderende sequenties worden "regulerende elementen" genoemd, en worden in vier klassen verdeeld naargelang hun functie: promotoren, versterkers ("enhancers"), isolatoren en verzwakkers ("silencers"). Onder hen spelen enhancers een cruciale rol in het activeren van genen als reactie op intra- en extracellulaire stimuli - hetgeen essentieel is voor de ontwikkeling van complexe organismen. Voorgaande onderzoeken suggereren dat het menselijk genoom meer dan een miljoen enhancers kan bevatten - en aantal veel groter dan het aantal eiwit-coderende genen (~22000). Er is echter niet veel bekend over de functie van de meeste enhancers, aangezien tot op heden slechts een klein aantal in detail onderzocht is.

In de afgelopen jaren heeft het gebruik van next-generation sequencing technologie (bijvoorbeeld RNA-seq) aangetoond dat de overgrote meerderheid van het menselijk genoom wordt getranscribeerd tot (niet-coderend) RNA. Aanvankelijk werd gedacht dat deze transcripten "junk RNA" van het "junk DNA" weerspiegelen. Deze hypothese werd later afgewezen door duizenden studies die aangeven dat deze niet-coderende RNA transcripten een opmerkelijk breed spectrum aan cellulaire processen regelen - waaronder transcriptie en translatie. Bovendien bleek dat de dysregulatie van regulerende elementen (en van de niet-coderende RNA transcripten die van hen worden getranscribeerd) in verband staat met verschillende menselijke pathologieën zoals kanker.In dit werk beschrijven we een gedetailleerd protocol van "Global Run-on sequencing (GRO-seq)", wat een high-throughput sequencing techniek is die net-onstane RNA transcripten meet. We hebben GRO-seq toegepast om enhancer-geassocieerde RNAs (eRNAs) te detecteren. Deze eRNAs zijn niet-coderende RNA transcripten, gemaakt door transcriptie van enhancer elementen. Onze experimenten hebben duizenden enhancers geïdentificeerd die geactiveerd worden door belangrijke transcriptie-factoren (bijvoorbeeld p53 en ER-alpha) en kunnen een rol spelen bij de ontwikkeling van kanker. On hun functie verder te karakteriseren hebben we een recent ontwikkelde technologie genaamd CRISPR-Cas9 gebruikt. Dit systeem bestaat uit een eiwit dat DNA kan splitsen (Cas9) en een nucleïnezuursequentie die het Cas9-eiwit kan leiden naar het doel (CRISPR). CRISPR-Cas9 bracht een revolutie in de biologie teweeg omdat het het mogelijk maakt om specifieke DNA-sequenties op een zeer snelle en eenvoudige manier te bewerken. Belangrijk is dat Cas9 kan worden toegepast op vrijwel alle sequenties in het menselijk genoom, waardoor het mogelijk is om de functie van enhancers en andere regulerende elementen op een uitgebreide manier te testen. Met CRISPR-Cas9 hebben we pionierswerk gedaan om de functie van enhancers te testen, door ze te muteren en het resulterende fenotype te onderzoeken met hoge doorloop (een zogenaamde genetische screen). Onze experimenten hebben geleid tot de identificatie van meerdere enhancers die de expressie van cruciale genen (bijvoorbeeld *CCND1*) en slecht gekarakteriseerde genen (bijvoorbeeld *CUEDC1*) in menselijke cellen regelen. In beide gevallen hebben we aangetoond dat deze enhancers absoluut noodzakelijk zijn voor de groei van kankercellen, en onze onderzoeksresultaten vormen de basis voor een betere diagnose en betere therapieën van kanker.

Curriculum vitae

Rui Filipe Marques Lopes was born in Mangualde da Serra (Portugal) on the 14th of November 1984. After finishing his high-school studies in Gouveia, he moved to Vila Real to study Biology at the University of Tras-os-Montes e Alto Douro. He concluded a major bachelor degree with distinction in 2008, after which he pursued a master in Oncology at the University of Porto. During this period he also worked as a research fellow at IPATIMUP (Portugal), where he studied the impact of germline mutations of E-cadherin in hereditary gastric cancer. In 2011, he was awarded with a PhD fellowship (FCT, Portugal) to carry on his doctoral studies at the Netherlands Cancer Institute in Amsterdam. In 2017, he joined the Novartis Institutes of Biomedical Research in Basel as a postdoctoral researcher. To date, he is co-author of nine scientific articles published in peer-reviewed journals.

List of publications

GRO-seq, a tool for identification of transcripts regulating gene expression. **Lopes R***, Agami R, Korkmaz G*. Methods Molecular Biology. 2017.

TGFβ1-induced leucine limitation uncovered by differential ribosome codon reading. Loayza-Puch F, Rooijers K, Zijlstra J, Moumbeini B, Zaal EA, Oude Vrielink JF, **Lopes R**, Ugalde AP, Berkers CR, Agami R. EMBO Reports. 2017.

Applying CRISPR–Cas9 tools to identify and characterize transcriptional enhancers. **Lopes R***, Korkmaz G*, Agami R. Nature Reviews Molecular Cell Biology. 2016.

Tumour-specific proline vulnerability uncovered by differential ribosome codon reading. Loayza-Puch F, Rooijers K, Buil LC, Zijlstra J, Oude Vrielink JF, **Lopes R**, Ugalde AP, van Breugel P, Hofland I, Wesseling J, van Tellingen O, Bex A, Agami R. Nature. 2016.

Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Korkmaz G*, **Lopes R***, Ugalde AP, Nevedomskaya E, Han R, Myacheva K, Zwart W, Elkon R, Agami R. Nature Biotechnology. 2016.

Myc coordinates transcription and translation to enhance transformation and suppress invasiveness. Elkon R, Loayza-Puch F, Korkmaz G, **Lopes R**, van Breugel PC, Bleijerveld OB, Altelaar AM, Wolf E, Lorenzin F, Eilers M, Agami R. EMBO Reports. 2015.

Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA. Léveillé N, Melo CA, Rooijers K, Díaz-Lagares A, Melo SA, Korkmaz G, **Lopes R**, Moqadam FA, Maia AR, Wijchers PJ, Geeven G, den Boer ML, Kalluri R, de Laat W, Esteller M, Agami R. Nature Communications. 2015.

p53 induces transcriptional and translational programs to suppress cell proliferation and growth. Loayza-Puch F, Drost J, Rooijers K, **Lopes R**, Elkon R, Agami R. Genome Biology. 2013.

E-cadherin destabilization accounts for the pathogenicity of missense mutations in hereditary diffuse gastric cancer. Simões-Correia J, Figueiredo J, **Lopes R**, Stricher F, Oliveira C, Serrano L, Seruca R. Plos One. 2012.

Acknowledgments

Disclaimer: this book is a product of many different contributors, so if you are reading this sentence it means you are one of them – thank you!

My adventure in Amsterdam lasted nearly 6 years. Like any adventure, it had ups and downs but if anyone asks me "*would you do it again?*", no doubt the answer is YES! At the professional level, it was an extraordinary experience and I learned much more than I ever anticipated. However, the personal experience was far more rewarding, and I will never forget it.

Reuven: thank you for taking the chance of applying with me for a PhD grant and giving me the opportunity to work in your lab. During this period I learned a lot from you and benefited from your mentoring in many aspects, from designing experiments to writing grants and papers. I was always impressed by your dedication to work and commitment to move the projects forward – I will never forget the first time you joined me in the lab to help doing a flow-cytometry experiment! Finally, thank you for trusting and giving me the freedom to explore new (and sometimes wild) ideas. After all, only high-risk projects have a chance to be published in Nature, Cell or Science – or worst case, Nature Biotechnology!

Gözde: I felt from the beginning that we had much more in common than just work, and Im very happy to see that this is still true more than 5 years later. We spent wonderful moments together, from lunch table to the lab, including seemingly never-ending experiments outside regular working hours (intermingled with videos of the one and only RiRi), and not forgetting dinners, parties and concerts (*Ruby, Ruby, Ruby, Rubyyy!*). I learned a lot from you, and you were always a role-model for me in the lab. I believe that our friendship and mutual trust was the key for success, and together we made it through troubled waters of cutting-edge science. Thank you!

I would also like to acknowledge the present and past members of the Agami lab for their input and criticism. Our lab meetings were always open for discussion and very interactive (once, I remember a speaker calling a voting on the veracity of the data that was on display). In particular, thank you Alejandro, Arieh and Pieter for your contributions to this work. Rani, thank you for the productive collaborations and being part of the doctoral committee. Ksenia, Lars and Sonia: it was very stimulating (and at times challenging) to guide you in the lab, but your enthusiasm and dedication to work was a constant. I was also fortunate to have collaborators outside the lab: thank you Ekaterina, Yongsoo and Wilbert for your effort in making this story a success.

Joost Gribnau and Ruud Delwel: thank you for accepting being part of the doctoral committee and evaluating my thesis. Fred van Leeuwen, Thijn Brummelkamp and Sjaak Neefjes: thank you for your feedback during the PhD committee meetings.

The NKI is a vibrant place to work, and during my PhD I was fortunate to meet outstanding people. Thank you to all the members of B4, H3 and H5 for the critical feedback about my work and for the good moments we spent at the NKI. In particular, the B4 meetings were very lively and helped me developing my critical sense. Mario, Tao and Bastiaan: thank you for inspiring discussions, sharing reagents, and troubleshooting. Heinz: thank you for your input during department meetings. I also appreciated our conversations around the coffee machine of H3. I would also like to thank Suzanne, Mariette and Thea for solving day-to-day problems and helping me settle in the Netherlands. Henry: you have many talents but the ones I appreciate the most are your ability to say the right words at the right time ("*when do you finish your PhD*?") and your sense of humor ("*you're still here after 6 years?!*"). Thank you for your support throughout my period at the NKI.

Fortunately, there was much more than work in Amsterdam, and my life was filled with many happy moments. Moments that made me feel welcome, like the weekly dinners in Louweshoek (Anna, Iris and Sannelien). Fortunately I moved out quickly, and then we could have dinners at my own place (Rita, Koos, Krystyna, Gözde, Andre, Mihoko, and Mario). One of the goals of my PhD was to be a "guest" member of B5, and for that reason I joined every single borrel they had or just showed up without any valid excuse (Ahmed, Melinda, Judith, Rita, Jonne, Mar, Femke, Lenno, Mihoko, Wytse, Indra, Jeroen, Iraia, Alba, Babet, Marion, Anna, Michiel, Daniela, Roy and Rene). When there was no borrel at B5, there was always the prospect of the Friday borrel (Tao, Eva, Laura, Christ, Joris, Roel, Santiago, Waseem, Michael, Andrea, Riccardo, Paula, Feline, Stefano, Chiara, Juliana, Aida, Cong, Jacqueline, Vincent, Joanna, Judith, Sedef, Loredana, Anirudh, Aurora, Inge, Judit, Zeliha, Jacobien, Lisanne, Heleen, Melanie, Chelsea, Marcelo, Eric, João, Mariana, Mathias, Isa!, Alessandra, Louise, Marjon, Ronak, Maria, Antonio, Simon, Ana Teresa, Ana Maia) or the department borrel (Heinz, Paul, Bas, Farshid, Muhammad, Hein, Bente, Thomas, Wietske, Tim, Stacey, Suzan, Laurel). After borreling, the priority was food and dinner was usually in the city center (Hayley, Marit, Dominika, Nicolas, Rubina, Carlos, Baastian, Tess, Jop and Mario). In-between experiments, there was time for coffee-breaks and to talk about anything but work (Alessandra, Telma, Rodrigo and Ferenc). There were nights in which I felt like a “victim” - usually at Pacific or KAH – but fortunately there were friends around (Rita, Anna, Lorenzo, Giusi, Mihoko and Arnold). During the short-lived Dutch summers, the Nieuw meer was a favorite for drinks and bites after work (Lorenzo, Renato, Alessandra, Petrit, Santiago, Jacobien, Giusi, Tao, Mario and Živa). And when Amsterdam felt too small, there was always someone ready to discover new places (Mario, Andrea, Pavel, Mihoko, Rita, João, Pedro, Cátia and Ângela). Sometimes, there was also work outside the lab, like painting walls, assembling furniture (Ahmed, Andre and Arnold). And when I decided to leave Amsterdam, it was time for disassembling and, yet again, I was not alone (João, Rita, Živa, Ahmed, Krystyna, Koos and Gözde). These memories only make sense because of you – thank you for being part of my life!

Koos: we shared the roller-coaster of doing a PhD and we made it! Our discussions were always stimulating and after talking to you it felt like anything was possible (like targeting enhancers with CRISPR-Cas9!). Your insights about biological questions are just amazing and I benefited tremendously from your critical comments. Thank you for listening and supporting when the future didn’t look so bright, and for celebrating the good moments as well! I wish you the best of luck for your dual-tenure as researcher… and daddy!

Krystyna: your energy, sense of humor and wittiness are contagious. Thank you for all the wonderful nights we spent around the dinner table! Moreover, thank you for caring and always having a kind word to share with me. I wish you all the best for the adventure of being a mom and I’m looking forward to meet your baby boy!

Živa: although you label yourself as a pessimistic, I always feel optimism when you are around. Thank you for listening, cheering up and giving the right advice. Good luck for the remaining of your PhD! PS Don’t forget to visit, otherwise I might show up in Slovenia. Or is it Slovakia?

Ahmed: I always enjoyed our meet-ups and football sessions. I was very happy to be by your side during your PhD defense and I hope to celebrate your future achievements. Otherwise, we will always have a reason to celebrate: Festivus, for the rest-of-us!

Andre: hey, buddy! Thank you for being such a great friend and always having a good word to share. Your sense of humor is just hilarious and I had lots of fun when you were around. I wish you success for your career in Tübingen and remember me when you need a “Director of Genomics”!

Rita: o teu nome está mencionado acima várias vezes, e isso não é por acaso. Durante cinco anos fomos não só colegas de casa, mas também camaradas de trabalho, companheiros de refeições, viciados em séries, “vítimas” da noite, sofredores do FC Porto (minuto 92!), e, acima de tudo, amigos. Obrigado por ouvires as minhas histórias (maioritariamente queixas acerca do trabalho), e teres sempre uma palavra de carinho e apoio. Desejo-te tudo de bom e muito sucesso para o futuro!

Cátia e Ângela: apesar de estarem longe, vocês estiveram sempre perto do coração. Bila forever!

João: provavelmente sabes tanto ou mais do que eu acerca desta tese. Não estou apenas a referir-me ao conteúdo científico, mas sim a tudo o que está por detrás deste livro. Durante estes seis anos, foram muitas as conversas acerca de trabalho, colegas de trabalho, chefes de trabalho, e acima de tudo acerca das nossas vidas. Também foram vários os *meetings*, em Mangualde da Serra, Vila Real e Amesterdão – sempre bem acompanhados pelo néctar dos deuses. As boas memórias são muitas, mas nunca me vou esquecer daquela viagem em que conduziste uma carrinha que mais parecia um camião! Obrigado por seres o Amigo com quem eu posso contar.

Por fim, dedico esta tese ás pessoas mais importantes da minha vida: António, Maria e Miguel. Pais, vocês serão sempre o meu exemplo de vida. É impossível expressar toda a gratidão que eu sinto, mas quero agradecer-vos pela educação que me deram e por todas as oportunidades que me proporcionaram ao longo da vida. Mano, tu estiveste do meu lado desde que eu tenho memória e eu sei que posso contar contigo para tudo. Eu apenas consegui chegar até onde estou graças à vossa generosidade e apoio incondicional. O vosso amor e carinho será sempre o melhor que eu tenho.