Development of a CRISPR-based epigenetic screening method

Anna Köferle

A thesis presented for the degree of

Doctor of Philosophy

University College London (UCL)

October 2016

Declaration

I, Anna Köferle, confirm that the work presented in this thesis is my own. Where information has been derived from other sources this has been indicated in the thesis.

_____

Anna Köferle

London, 20 October 2016

Abstract

Millions of chromatin marks have been profiled across the human genome in different tissues and cell types. Yet, which and how many of these marks contribute to the establishment and control of gene activities remains incompletely understood. The focus of this PhD project is to develop a CRISPR-based epigenetic screening method for the discovery of functional epigenetic marks. The aim of this method is to identify sites in the genome where addition or removal of a particular chromatin mark has an impact on cellular phenotype. To this end, I fused the catalytic domain of a chromatin-modifying enzyme to the nuclease-dead Cas9 protein and introduced it into cells concomitantly with an appropriate library of guide RNAs (gRNAs). Cells that show a change in the phenotype of interest are then separated from the pool of cells by Fluorescence-activated cell sorting (FACS). I designed libraries of gRNAs that are both targeted towards a particular phenotype of interest, targeting regulatory regions around a limited set of genes, as well as more complex, genome-wide libraries, which were generated from fragmented genomic DNA. I used the CRISPR-based screening strategy to identify candidate gRNAs, which, together with the appropriate dCas9-chromatin modifier fusion protein, bring about changes in expression of various cell surface markers.

Acknowledgement

I would like to thank my supervisor Stephan Beck for giving me the opportunity to work in his lab, for being supportive and incredibly patient and giving me the freedom to pursue my own ideas by trial and error.

I would also like to thank Stefan Stricker for being a fantastic collaborator, always full of ideas and eager to discuss experiments.

Furthermore, I would like to acknowledge all the members of the Beck lab past and present for creating a wonderful and friendly work environment. In particular, I would like to thank Andrew Feber for his advice and help throughout the past three years, for helping everyone who has ever worked in this lab with countless little things - here is a big Thank You! I would also like to acknowledge Gareth Wilson for helping me with my first foray into bioinformatics, and James Barrett for help with statistics and data analysis issues. I learnt a lot from both of you. Pawan Dhami, Paul Guilhamon, Dirk Paul, Sabrina Stewart, Miljana Tanic, Yuan Tian, Amy Webster - you have been the kindest, funniest, nicest people to work with and I would not have made it through this PhD without your support, or at least it wouldn't have been as much fun!

I would further like to acknowledge my fantastic collaborators who are based in Munich: Karolina Worf, Christiane Fuchs, Christopher Breunig and Valentin Baumann.

Finally, I would like to thank my parents for everything they have done for me and Luki, for always being there for me, providing help and moral support and keeping me sane. I couldn't have done this without you!

Publications

The following publications resulted from work conducted during the course of this PhD project:

Stricker, S. H., **Köferle, A.**, Beck, S. (2017). From Profiles to Function in Epigenomics. *Nature Reviews Genetics* 18:1, 51–66. doi:10.1038/nrg.2016.138

Köferle, A., Worf, K., Breunig, C., *et al.* (2016). CORALINA: A universal method for the generation of gRNA libraries for CRISPR-based screening. *BMC Genomics*, 1–13. http://doi.org/10.1186/s12864-016-3268-z

Köferle, A., Stricker, S. H., Beck, S. (2015). Brave new epigenomes: the dawn of epigenetic engineering. *Genome Medicine*, 1–3. http://doi.org/10.1186/s13073-015-0185-8

This article was selected by the Scientist magazine to be adapted for an online opinion article `http://www.the-scientist.com/?articles.view/articleNo/43837/title/Opinion--Engineering-the-Epigenome`

Wilson, G. A., Lechner, M., **Köferle, A.**, *et al.* (2013). Integrated virus-host methylome analysis in head and neck squamous cell carcinoma. *Epigenetics*, 8:9, 953–961. http://doi.org/10.4161/epi.25614 (rotation project)

Contents

List of Figures

List of Tables

List of Abbreviations

A549	Lung adenocarcinoma cell line
ChIP-seq	Chromatin immunoprecipitation followed by next-generation sequencing
CRISPR/Cas	Clustered regularly interspaced short palindromic repeats/CRISPR-associated
crRNA	CRISPR RNA
CTCF	CCCTC-binding factor
dCas9	Nuclease-dead version of CRIPSR associated protein Cas9
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease enzyme
DNMT	DNA methyltransferase
EMT	Epithelial-to-mesenchymal transition
ENCODE	Encyclopedia of DNA Elements Consortium
ES cells	Embryonic stem cells
FACS	Fluorescence-activated cell sorting
FOXA1	Forkhead box protein A1
G9a	Histone methyltransferase, also EHMT2 -Euchromatic histone-lysine N-methyltransferase 2
GATA1	GATA-Binding Protein 1 (Globin Transcription Factor 1)
GFP	Green fluorescent protein
gRNA	Guide RNA
HAT	Histone acetyltransferase
HbF	Foetal haemoglobin
HDAC	Histone deacetylase
HEK293T	Human embryonic kidney cell line

HP1	Heterochromatin Protein 1
HSF1	Heat shock transcription factor 1
IF	Immunofluorescence
IHEC	International Human Epigenome Consortium
JMJD2A	Jumonji Domain-Containing Protein 2A, also KDM4A - Lysine Demethylase 4A
K562	Chronic myelogenous leukemia cell line
KLF1	Krüppel-Like Factor 1 (erythroid)
KRAB	Krüppel associated box, transcriptional repressor
LCR	Locus Control Region
LSD1	Lysine Specific Demthylase 1, H3K4me and H3K9me specific demethylase
MET	Mesenchymal-to-epithelial transition
MNase	Micrococcal nuclease
MOI	Multiplicity of infection
mRNA	Messenger RNA
MYOD1	Myoblast determination protein 1
NFκB	Nuclear factor κB
NHEJ	Non-homologous end joining
PAGE	Polyacrylamide gel electrophoresis
PAM	Protospacer adjacent motif
PBMCs	Peripheral Blood Mononuclear Cell
PRDM9	PR domain zinc finger protein 9, H3K4 trimethyltransferase
RCC4	Clear cell renal cell carcinoma cell line
RNA	Ribonucleic acid
RNAi/shRNA	RNA interference/ short hairpin RNA
RNAPII	RNA polymerase II enzyme
SET7	SET Domain Containing Lysine Methyltransferase 7
SETD2	SET Domain-Containing Protein 2, H3K36 methyltransferase
SUV39H1	Suppressor Of Variegation 3-9 Homolog 1, a histone H3 Lysine-9 specific methyltransferase

SWI2/SNF2	SWItch/Sucrose Non-Fermentable, a nucleosome remodelling complex
TAD	Topologically-associating domains
TAFII250	Transcription initiation factor TFIID 250 kDa subunit
TALE	Transcription activator-like effector
TCF3	Transcription factor 3
TET	Ten-eleven translocation methylcytosine dioxygenase 1
TF	Transcription factor
TFIID	Transcription factor II D
TGF-β	Tumour growth factor β
TNF-α	Tumour necrosis factor α
tracrRNA	Transactivating CRISPR RNA
TSS	Transcriptional start site
UMI	Unique molecular identifier
VP64	Transcriptional activator protein
ZF	Zinc finger protein

Chapter 1

Introduction

1.1 Epigenetic gene regulation

Non-genetic factors contribute to many cellular functions and phenotypes [1]. Among the first to recognise this was C. H. Waddington, who introduced the term "epigenetics" in 1942 to describe molecular mechanisms through which "the genes of the genotype bring about phenotypic effects" [2]. Epigenetics is thus defined as the study of regulation of gene activity (although many other definitions of the term are also commonly used).

One of the first steps at which a gene's activity is regulated is at initiation of transcription, effectively deciding whether the information stored in a gene is read out or not. Transcription is the copying of information stored in one strand of template deoxyribonucleic acid (DNA) into ribonucleic acid (RNA) by the enzyme RNA polymerase. Several mechanisms converge to regulate this first step of gene expression: The local binding of transcription factors and polymerases to promoters, chromatin marks on the nucleosomes associated with DNA and the DNA itself, and potentially, global positioning of genes within the nucleus all influence whether transcription is initiated. Each of these regulatory mechanisms will be introduced briefly below before examining, in greater detail, the role of chromatin marks in gene regulation, which is the focus of the work conducted for this PhD thesis.

1.1.1 Transcription factors

Regulation of transcription at the promoter occurs through site-specific binding of transcription factors (e.g. sigma factors) that recruit RNA polymerase [3]. Transcription factors (TFs) are proteins that recognise DNA sequences that are usually 5-15 bp in length and that often bind directly to gene promoters or to cis-regulatory elements such as enhancers and together with other "co-activators" and "co-repressors" regulate assembly of the transcriptional machinery at the promoter.

Transcription factor activity can be regulated by post-transcriptional modification, synthesis and degradation and through sequestration by or release from regulatory proteins. For example, during heat shock, the conserved heat shock transcription factor 1 (HSF1) induces the expression of heat shock genes that protect cells against a number of external stresses and assist in the repair of damaged proteins. HSF1 is present in an inactive state under normal conditions and is acti-

vated rapidly by heat stress. Activation of HSF1 is accompanied by trimerization and high-affinity binding to highly conserved DNA regions known as heat shock response elements in the promoters of heat shock genes [4]. Another example of a transcription factor that activates genes rapidly in response to an external signal is NFκB [5]. It is normally kept in an inactive state through inhibition by regulatory proteins (IκB). Pro-inflammatory cytokines such as tumor necrosis factor α (TNF-α) trigger intracellular signalling pathways that lead to phosphorylation and ubiquitinylation of the inhibitor, which is subsequently degraded by the proteasome, thus freeing NFκB from inhibition. This signalling mechanism leads to activation of NFκB-responsive genes within minutes in response to TNFα.

The number of transcription factors encoded in the human genome is estimated to be 1,000-3,000 [6]. Given this large diversity and the number of binding sites found in the genome that may be bound or not bound at any given moment, as well as potential for synergistic effects [7], the regulatory network formed by transcription factors is complex and difficult to disentangle. The genome-wide occupancy of a particular transcription factor can be assayed using methods such as ChIP-seq, whereby transcription factors and bound DNA are cross-linked and an antibody against the transcription factor is used to pull down and enrich for bound DNA sequences. These DNA segments can then be sequenced to map the genome-wide binding profile of the transcription factor and also to define its consensus binding site. Transcription factor occupancy has been profiled for many different factors in various organisms, cell types, and different experimental conditions. The emerging picture of the regulatory network formed by transcription factors is complex. In myoblasts at different stages of differentiation, the transcription factor MYOD1 (myoblast determination protein 1) is found to bind to some sites regardless of differentiation stage, but to some exclusively at a particular stage [8]. A similar pattern of universally bound sites as well as developmental stage-specific binding events have been identified for the transcription factor TCF3 in B-cell specification [9] and KLF1 during erythrocyte differentiation [10] amongst many other examples. In many cases it is not clear how this differential binding is regulated and while interesting, the observed differences in occupancy do not necessarily imply a causal effect on gene expression. When TF occupancy for a large set of transcription factors was integrated with a large library of expression data from yeast obtained under different conditions, around 50 % of binding events could be correlated with changes in expression [11]. In mammalian cells this correlation was lower, with a 10-25 % overlap [12]. It thus appears that changes in binding of a single transcription factor may not always be sufficient

to elicit a transcriptional change. Binding could also simply be a consequence of the DNA being accessible. Some transcription factors also act as pioneer factors and prime an inactive locus for later transcription, e.g. by remodelling chromatin and changing nucleosome positioning, but do not themselves activate transcription. How long a transcription factor stays bound at the promoter may also be important. Indeed, some studies have suggested that rather than profiling occupancy using methods such as ChIP-seq, measuring the residence time of a transcription factor at a particular site might be a better indicator of function [13].

1.1.2 Chromatin modifications

One of the factors that may influence whether a transcription factor can efficiently bind its target binding site is accessibility. In eukaryotes, DNA exists in the form of chromatin and is wrapped around histone proteins to form nucleosomes. Chromatin function and accessibility is thought to be regulated by numerous chromatin modifying enzymes, which add chemical groups to or remove them from the tails and core residues [14] of histone proteins or DNA bases (**Figure 1.1**). Chromatin remodellers reshuffle entire nucleosomes, the basic unit of chromatin in which 147 bp of DNA are wrapped around one histone octamer consisting of two copies of the core histones H2A, H2B, H3 and H4 and the linker histone H1 [15]. It is thought that chromatin modifications influence virtually all processes taking place on chromatin including transcription, replication and DNA repair.

Several models have been proposed for how histone modifications function: The **"histone code" hypothesis** [16] (**Figure 1.2 A**) states that chromatin marks act in a combinatorial (or sequential) fashion to specify a particular functional output. Chromatin modification is further thought to impact gene regulation through two principal modes of action: (1) by affecting the physical properties of chromatin and regulating access to binding sites and (2) by providing a platform for signal transduction. For example, acetylation of positively charged lysines in histone tails destabilises the association with negatively charged DNA. This does not compromise the integrity of the nucleosome but might contribute to providing easier access to transcription factors and polymerases (**"charge neutralization model"**, **Figure 1.2 B**) [17]. In accordance, lysine acetylation correlates with active transcription.

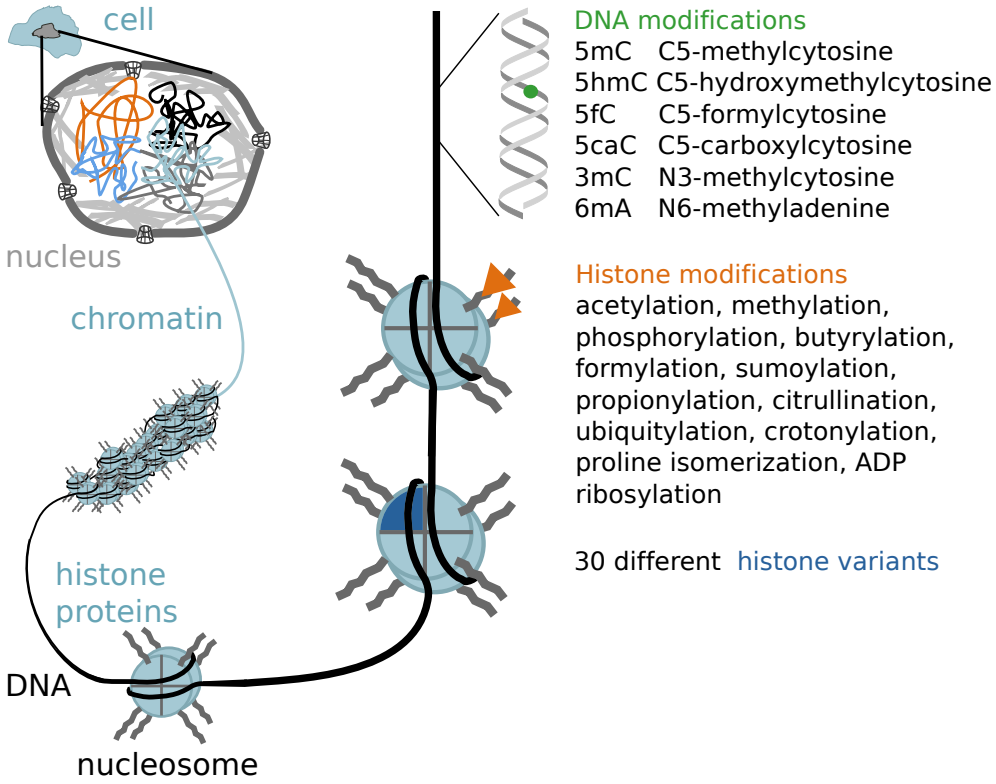


Figure 1.1: Chromatin and its modifications. The cytosine and adenine bases of DNA as well as the N-terminal tails and core residues of histone proteins have been found to be chemically modified as shown. Entire histones may be exchanged with variant histones, e.g. H3.3 often replaces H3 in nucleosomes at active genes and γ-H2AX is incorporated at sites of DNA repair of double-strand breaks.

The **"signalling network model"** of chromatin [18] emphasizes that post-translational modifications of histones create docking sites for regulatory proteins (**Figure 1.2 C**). For example, in addition to potentially affecting the interaction between histones and DNA, acetylated lysines in histone tails can be recognised by bromodomain-containing proteins including TAFII250, which is part of the general transcription factor complex TFIID, a core component of the pre-initiation complex that forms at promoters upon initiation of transcription, and by SWI2/SNF2, a nucleosome remodelling complex. The signalling model further postulates that multiple modifications can combine to confer network properties of bistability (switch-like behaviour due to presence of feedback loops and thresholds) and robustness (due to redundancy). A well-characterised positive feedback loop is the spreading of silencing H3K9 methylation mark, which is deposited by SUV39H1, then recognised by the chromodomain-containing protein HP1, which in turn recruits more SUV39H1 methyltransferase [19, 20].

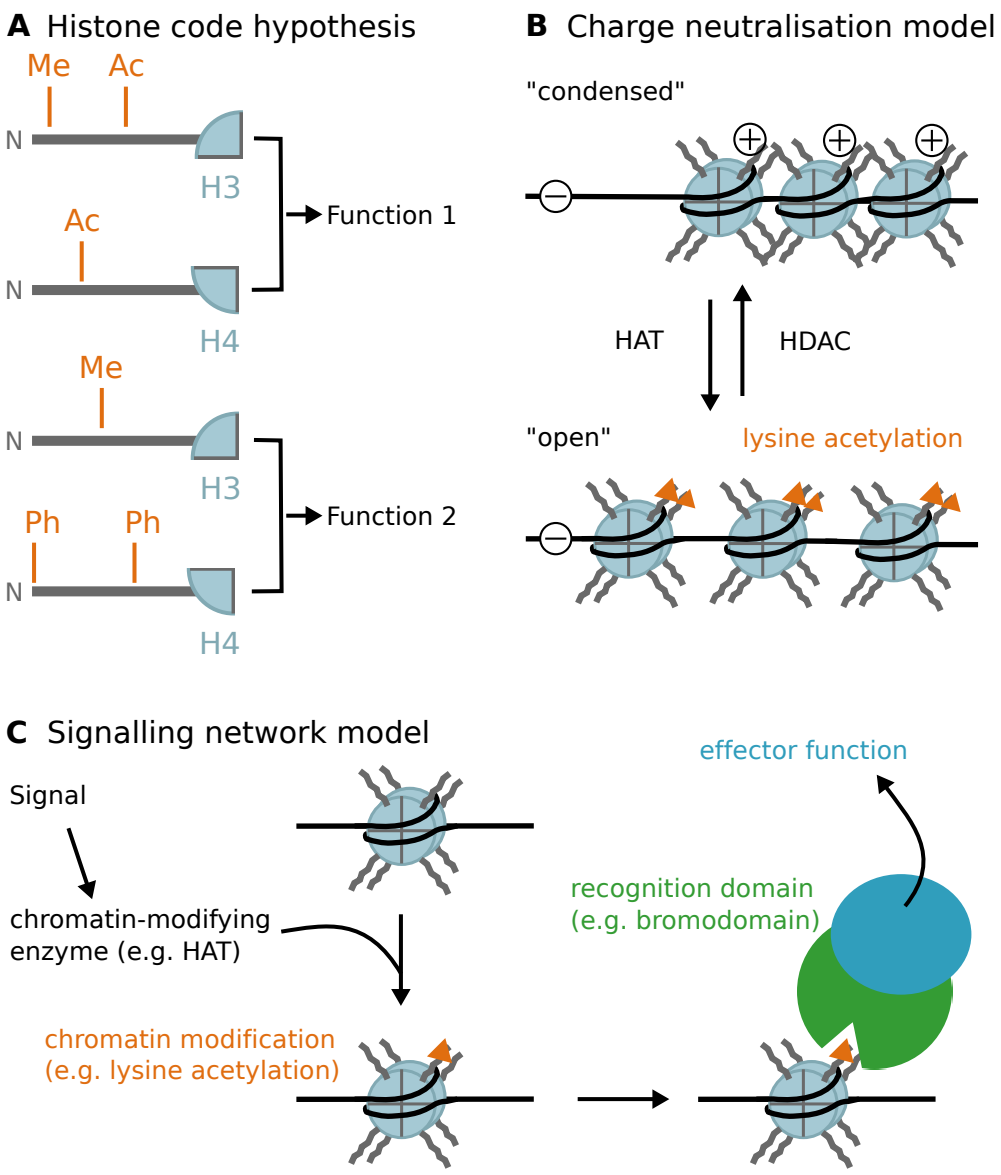


Figure 1.2: Models for chromatin function. A. The histone code hypothesis proposes that chromatin marks act in a combinatorial (or sequential) manner to specify a particular functional output. **B.** The charge neutralisation model describes how addition or removal of charges from the N-terminal tails of histones may affect interactions between histones and negatively charged DNA. **C.** The signalling network model envisions chromatin modification as part of a signal transduction pathway, whereby chromatin marks are added or removed in response to an upstream signal and then form a platform for binding of effector proteins.

Like transcription factor binding sites, chromatin modifications have been mapped genome-wide by ChIP-seq in a wide variety of tissues and cell types using antibodies specific to particular post-translational modifications on N-terminal histone tails. Profiling of a total of 37 different histone marks in $CD4^+$ T cells [21, 22] has revealed that many gene promoters have unique combinations of marks. However,

in concordance with previous studies [23] including those employing ChIP-qPCR [24] or ChIP-chip [25] methods, it was found that the promoter regions of known active genes tend to have elevated histone acetylation as well as high levels of H3K4 methylation (mono, di and tri) and H3K36me3 is present over the gene body. Active genes also show reduced nucleosome occupancy around the transcriptional start site (TSS) [26]. Repressed genes, on the other hand, were found to harbour trimethylation of H3K27, H3K9 and H3K79 (**Figure 1.3**).

A quantitative model based on the genome-wide histone modification data from $CD4^+$ T cells aimed to interrogate the relationship between histone modification level and gene expression and achieved good correlation when comparing modeled and measured expression levels (Pearson correlation coefficient r= 0.77 when comparing to microarray expression data, r = 0.81 when comparing to RNA-seq data) [27]. It further became evident that a large number of marks correlate with each other, e.g. H3K27ac and H2BK5ac levels had a correlation coefficient r = 0.97, so in principle it is sufficient to profile a subset of known marks to infer the transcriptional status of a locus. A model based on only three marks (H3K27ac, H3K4me1, H4K20me1) still predicted expression well (r max = 0.75).

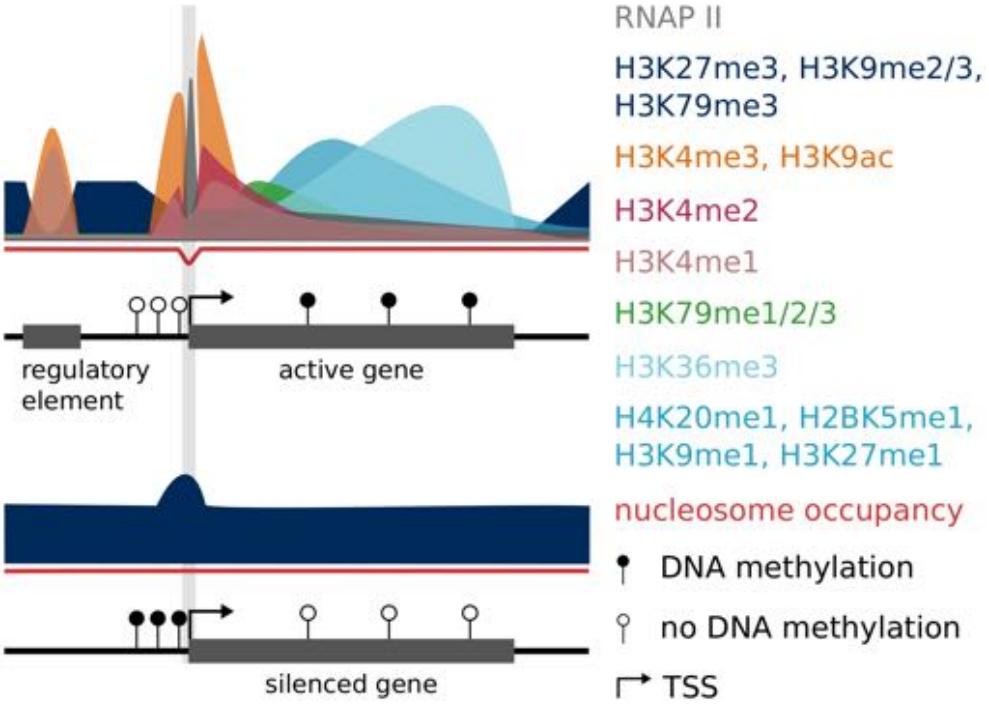


Figure 1.3: The chromatin landscape around active and inactive genes (adapted from [21] and [28]).

Genome-wide profiles of many different chromatin marks are now available for a large number of cell types and experimental conditions, in part due to large international efforts, such as IHEC (International Human Epigenome Consortium) [29], ENCODE (Encyclopedia of DNA elements) [30] and NIH Roadmap Epigenomics Mapping Consortium [31]. ENCODE originally mapped the location of up to 12 histone modifications and histone variants in 46 different cell types (as well as generating ChIP-seq profiles for 119 different DNA-binding proteins and a number of RNA polymerase components in 72 cell types) [32]. At the time of writing, the IHEC data portal (`http://epigenomesportal.ca/ihec/`) contained 7,132 datasets referenced to the human genome, out of which the majority (3,725) were profiles of histone modifications. A more in-depth discussion of what can be learned about the function of chromatin marks from comparing these profiles as well as a discussion of approaches to identify regulatory chromatin marks can be found in section 1.2.

1.1.3 Nuclear architecture

The molecular machinery that is involved in transcription as well as its regulation has been found to be non-uniformly distributed in the cell's nucleus [33]. Chromosomes appear to be confined to regions of the nucleus called chromosome territories [34]. The folding of chromatin in three-dimensional space has been studied in a variety of organisms and cell types and overall chromatin architecture appears to be remarkably conserved between several mammalian genomes [35]. Chromatin is further organised into so-called "topologically-associating domains" (TADs), where sections of chromatin interact much more frequently with each other than with segments that lie outside the domain.

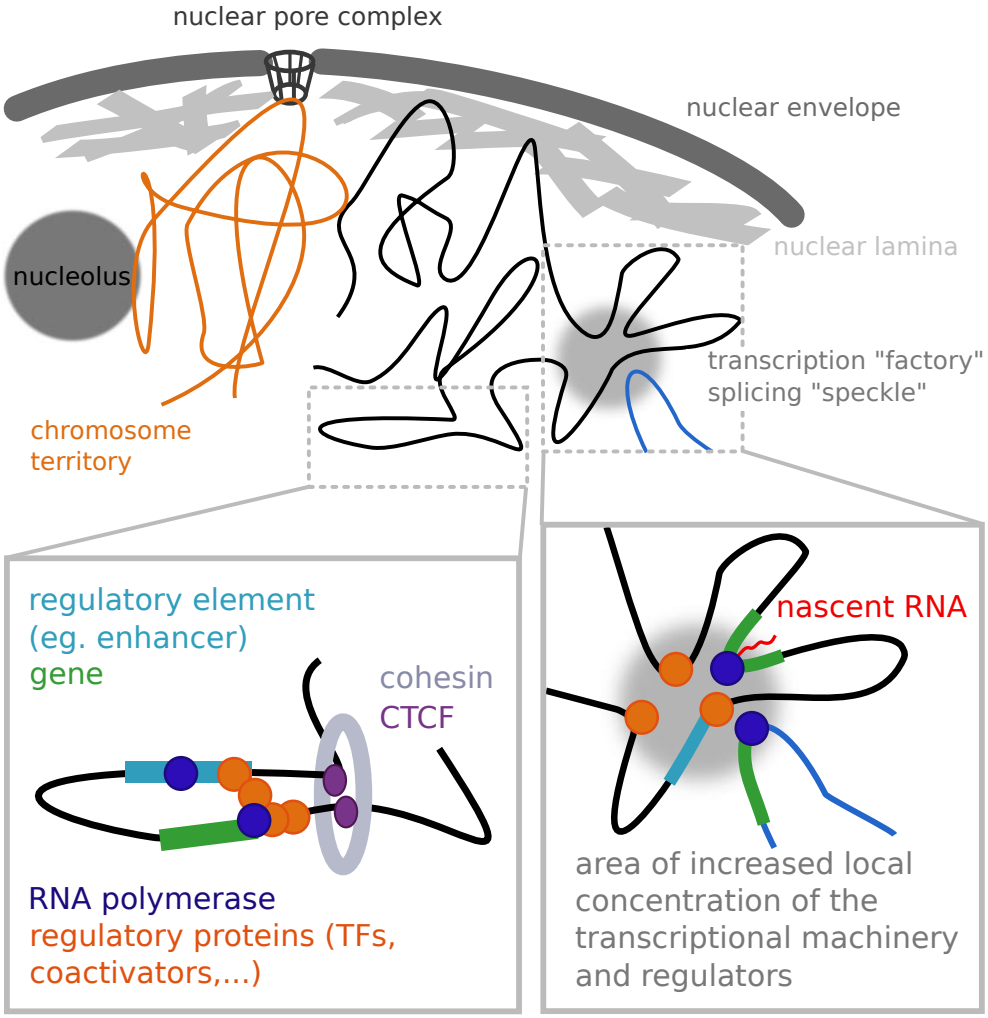


Figure 1.4: Transcription in the context of nuclear architecture. Chromatin, containing approximately 2 m of DNA for a human genome, is folded in 3D space to fit into the nucleus which is approximately 6 μm in diameter. Regions of chromatin that interact with the nuclear lamina or are located at the nuclear periphery appear to be condensed and inactive ("heterochromatic"). Transcription (as well as subsequent RNA processing) appears to take place in foci with a high local concentration of the transcriptional machinery and RNA processing machinery. These structures are dynamically assembled and have been termed transcription "factories" and splicing "speckles" respectively. The nucleolus is a specialised site for transcription of ribosomal genes by RNA polymerase I and III. Chromatin loops can be formed transiently (through the act of transcription in a factory) or be maintained by scaffolding proteins. Chromatin looping is also thought to be the mechanism through which enhancers interact with their target promoters.

These TADs can be several kilobases to megabases in length and domain boundaries are defined by insulator regions and CTCF binding sites in particular orientation. Chromatin loops contained within TADs often link enhancers and promoters in three-dimensional space. Enhancers are defined as DNA sequences

that can increase transcription of a target gene, usually in *cis*. They can do so from considerable distances ranging from as close as 100 bp to megabases away [36]. Long-range interactions between enhancers and promoters, which may be far apart on the linear genome but may be brought into close spatial proximity through chromatin looping, are important for gene regulation (**Figure 1.4**, left inset)[37]. At the beta-globin locus for example, interactions between the locus control region (LCR) and downstream regulatory elements drive the formation of a 200 kb loop, specifically in cells expressing the locus [38]. Genetic inversion of CTCF binding sites that define domain boundaries can alter domain topology and has been shown to impact negatively on transcription when enhancer-promoter interaction are disrupted [39].

Interchromosomal contacts are much more infrequent than contacts between regions within a TAD, but they have been observed at several well-studied loci during transcription of co-regulated genes (**Figure 1.4** right inset) [40–45]. Whether observed changes to nuclear architecture are purely a consequence of transcription or whether they can also regulate gene expression remains difficult to establish. While it is clear that transcription occurs in the context of nuclear architecture and also shapes it, the notion that a cell uses nuclear architecture to regulate gene expression is still under debate.

1.1.4 Tug of war between different modes of gene regulation

A plethora of signals - both extracellular and intracellular - need to be integrated at the promoter of a given gene to specify a transcriptional output. Inevitably, different mechanisms of gene regulation converge at this point. One can think about the integration of activatory and inhibitory signals as a tug of war. For example, the chromatin environment might be favourable for transcription but without the required transcription factors being expressed in that cell, transcription will not be initiated. Conversely, an active transcription factor might not be able to initiate transcription if its binding site at a promoter or enhancer is inaccessible. However, some pioneer transcription factors can themselves change chromatin environment through recruitment of other chromatin-modifying or chromatin-remodelling proteins. There is thus a dynamic interplay between different modes of gene regulation. The transcription factor FOXA1, for example, acts as a pioneer factor involved in the expression of liver-specific genes during development and is also involved in the recruitment of Estrogen Receptor in breast cancer cells,

as well as androgen receptor recruitment in prostate cancer cells. *In vitro* reconstitution experiments have demonstrated that FOXA1 can displace nucleosomes and create so-called DNase-hypersensitive sites even without the recruitment of additional chromatin remodellers [46]. (This is thought to be due to structural similarity to the linker histone H1, which could enable FOXA1 to physically disrupt the interaction between DNA and histones.) However, another study also found that FOXA1 is predominantly present at enhancers rich in H3K4me1/me2 but poor in H3K9me2 [47]. This correlation is consistent with the notion that chromatin context dictates where FOXA1 binds to chromatin.

The extent to which chromatin context or transcription factors dominate at a given promoter probably has to be established on a case-by-case basis. It has been shown in *D. melanogaster* and *C.elegans* that several developmentally responsive genes are transcribed despite being completely devoid of the "activating" chromatin marks that were illustrated in **Figure 1.3** above [48]. It is possible that rapid activation and inactivation of a class of genes at a particular time point in development is entirely dependent on external signals with chromatin context playing an insignificant role.

The next section will focus on how big a contribution chromatin marks can make towards regulation of transcription and whether alteration of individual chromatin can force a gene into an active or inactive state respectively.

1.2 Can changes in chromatin marks regulate gene expression?

While the presence or absence of certain chromatin marks often correlates with transcriptional activity at a locus, it is difficult to establish whether changes in chromatin marks can actually cause changes in gene expression as correlation does not necessarily imply causation. In this section, attempts to infer function from profiles of chromatin marks will be briefly introduced, before discussing the use of genetic approaches to provide evidence for the functional role of chromatin marks in gene regulation. Such approaches can be broadly separated into the following classes: genetic manipulation of the DNA sequence underlying a chromatin mark or larger chromatin feature, genetic mutation of histone proteins, and mutation of the enzymes that add or remove chromatin marks. This will be followed by a section describing novel approaches for editing individual chromatin marks.

1.2.1 Inferring function from profiles of chromatin marks

Profiling of chromatin marks is an important first step towards identifying associations between chromatin features and genomic function at the level of gene regulation. The genome-wide distributions of different chromatin marks have been catalogued for different tissues and cell types and are remarkably useful for predicting transcriptional rates [49]. A set of chromatin marks has also been used to computationally impute chromatin states [50]. When integrated with mapping of DNase hypersensitive sites, annotation of transcriptional start sites, transcripts and exons, as well as genome-wide binding profiles of CTCF, c-MYC, and NF-κB, a subset of these "learned" states was found to identify "promoter", "enhancer", "insulator", "transcribed" or "repressed" regions of the genome. A chromatin state is thus an annotation of an inferred function, based on the observed combination of epigenetic marks at a particular location in the genome. Several of the inferred enhancers, defined by strong H3K4 methylation and weak RNAPII signals, have also been experimentally validated [50].

Profiles of marks generated in different cell types or from the same source over time can also be compared to identify regions that vary between different conditions. Such a "comparative" approach has been used to find regions that differ between embryonic stem cells and differentiated cell types in mouse [51]. This led to the discovery that embryonic stem cells possess bivalent domains, which harbour both "activatory" and "repressive" chromatin marks, in the promoters of genes important for development. While not able to prove function, such approaches can highlight interesting candidate regions for further experimental analysis.

1.2.2 Insights from mutational analysis of chromatin-modifying enzymes

It is possible to mutate or remove single bases harbouring DNA modifications. However, this approach is not applicable to individual histone modifications and can only be used to remove marks, not to add them. Nevertheless, such approaches have been useful in linking several epigenetic mechanisms including DNA methylation [52], chromatin looping [53] and noncoding transcription [54] to genomic imprinting. However, genetic manipulation can only provide indirect evidence for causality. In most cases, entire genomic domains containing the feature of interest were excised by gene targeting in these studies, which makes it

impossible to disentangle the effect of loss of a mark from the effect of loss of the underlying DNA sequence.

It has been established that a large number of chromatin modifying enzymes are essential for normal development and their loss induces embryonic lethality, in some cases relatively early in mouse development (for example *Dnmt1* [55], *Dnmt3a* and *Dnmt3b* [56], *Lsd1* [57], *Hdac1* [58], *Suv39h* [59]). However, embryonic development is complex and can be disturbed in many ways, in fact there are currently 2,669 mouse genes with embryonic lethality annotation [60]. Furthermore, when chromatin-modifying enzymes or their histone targets are mutated, marks become globally altered across the genome. For example, overexpression of mutant H3.3K27M in human neural progenitor cells leads to a global reduction in histone H3K27me3 and is accompanied by the induction of genes associated with a less differentiated developmental stage [61]. Conditional deletion of the H4K20 dimethyltransferase SUV4-20h1 in skeletal muscle cells in mice causes global reduction of H4K20me2, a concomitant increase in H4K20me1 and a reduction in H3K27me3 levels, while H4K20me3 levels remain unchanged. Mutant mice show decreased amounts of heterochromatin in the nuclei of skeletal muscle stem cells, increased levels of *MyoD* expression (20-fold) and exhaustion of stem cells upon repeated injury leading to a muscle regeneration defect [62]. Given that marks are altered globally in these experiments, they cannot provide insight into the function of chromatin marks at individual sites. In addition, care has to be taken when interpreting the results of chromatin modifier knockouts with respect to transcription. Most, if not all, chromatin-modifying enzymes have non-histone targets as well [63–65]. Therefore, phenotypic changes resulting from chromatin modifier knockouts (or pharmacological inhibition) cannot be attributed to misregulation of chromatin alone.

One important insight gained from knockout studies is that global loss of chromatin marks does not always lead to major changes in the transcriptome. Surprisingly limited transcriptional changes were detected following the almost complete loss of DNA methylation through triple knockout of *Dnmt1*, *Dnmt3a* and *Dnmt3b* in mouse ES cells [66]. Following knockdown of SETD2 in RCC4 renal cancer cells, ChIP-seq revealed loss of H3K36me3 along 2,513 genes, but no global changes in RNAPII binding could be detected and only 326 genes showed changes in transcription detectable by RNA-seq [67]. It is possible that a small but functionally relevant fraction of marks escaped removal in these studies. However, it is also possible that a majority of chromatin marks do not play functional role with respect to transcription, at least in stable cell populations in culture. Perhaps

the functional importance of marks would only be revealed by major transitions such as during differentiation, reprogramming or transformation.

1.2.3 Recent advances in genome editing methods

It has recently become possible to design DNA binding proteins using either a Zinc-finger (ZF), TALE (Transcription activator-like effector), or CRISPR/Cas (Clustered regularly interspaced short palindromic repeats/CRISPR-associated) architecture [68]. These proteins can be engineered to bind to a unique user-defined site in the genome. With Zinc-fingers and TALEs, targeting to a specific genomic sequence is achieved by a programmable DNA binding domain. This domain is constructed by combining modular protein domains that each recognise a particular base triplet (ZF) or single base (TALEs) in the target sequence. The bacterial protein Cas9 on the other hand can be targeted by a synthetic RNA molecule, which is a fusion of the so-called crRNA and tracrRNA of the type II CRISPR system of *Streptococcus pyogenes* [69–71]. This guide RNA (gRNA) consists of a 20 bp protospacer sequence that determines the target binding site and a scaffold sequence that folds into a stem loop which is recognised by the CRISPR-associated protein Cas9. Targeting with CRISPR is straightforward compared to other platforms because the targeting sequence can be used directly as a template for the gRNA sequence. The only requirement for gRNA design is the presence of a protospacer-adjacent motif (PAM) sequence of the form "NGG" (for this particular Cas9, see [72]) immediately after the targeting site. Using a protein-based targeting system, each base in the targeting sequence has to be matched with the corresponding protein domain. Design of Zinc finger is complicated by crosstalk between adjacent protein domains in the targeting moiety. TALE assembly is challenging because it involves a multi-step cloning procedure to juxtapose repetitive domains as the base specificity is only conferred by two amino acids within each repeat region.

All three systems have been used successfully for genome editing. For genome editing purposes, the DNA double strand has to be physically broken in order to trigger repair via non-homologous end joining (NHEJ), resulting in small insertions or deletions, or Homology-directed repair, leading to precise changes (insertion/deletion of a defined sequence or single nucleotide changes) through use of a plasmid-based repair template. In order to trigger formation of a double-strand break at a precise genomic location, the DNA binding domain of Zinc fingers and TALEs is fused to a nuclease or nickase domain. Cas9 on the other hand has

endogenous nuclease activity. These novel methods have made it much faster and easier to produce genome modifications compared to earlier gene targeting methods and they have already been used generate knockouts of chromatin modifiers [66]. However, these experiments suffer from the same caveats as the experiments described above, primarily in that they cannot distinguish the relative contributions that loss of the chromatin modification and loss of the underlying DNA sequence make to the observed effect.

1.2.4 Epigenome editing with programmable chromatin modifiers

Chromatin modifying enzymes, or their minimal catalytic domains, can be made targetable through fusion to a zinc-finger or TALE DNA binding domain or to the catalytically inactive version of the CRISPR protein Cas9 (dCas9). Knockout of the nuclease activity of Cas9 is achieved by introducing two single base pair changes (D10A and H840A) [69]. These proteins act as a targeting platform, changing individual (or a few) chromatin marks at a specific site in chromatin without altering the underlying DNA sequence.

A number of chromatin-modifying enzymes have already been attached to different DNA binding domains. These have been used to add or remove chromatin marks at the target sites (see **Tables 1.1** and **1.2** for details of these studies and observed effects, see also [73]). Collectively, these studies have shown that catalytic domains of chromatin-modifying enzymes are sufficient to induce transcriptional changes when directed to specific target sites. Adequate controls were used in most of these studies, including catalytic mutants which ensured that the observed effect is due to enzymatic activity and not merely due to chromatin binding as well as off-target controls to control for overexpression of the chromatin modifier.

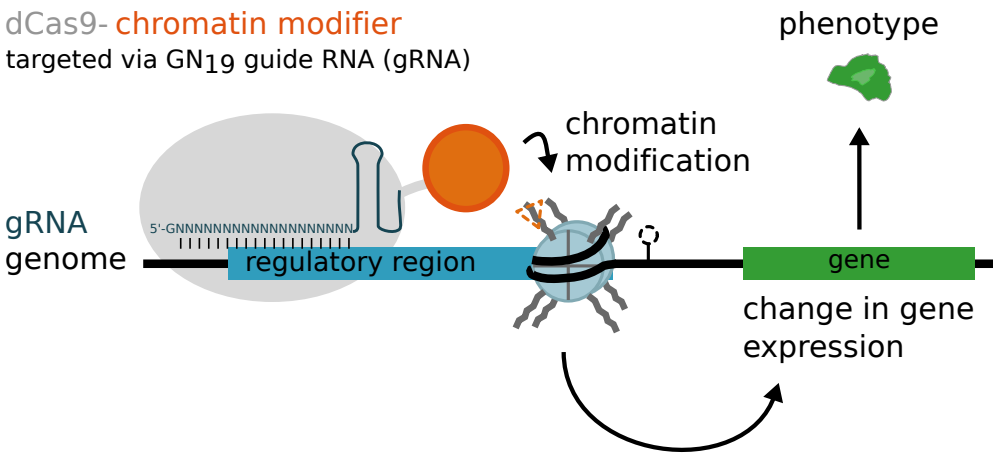


Figure 1.5: Epigenome editing with the CRISPR/Cas system. A synthetic guide RNA (gRNA) can be used to target an effector (orange), fused to the catalytically dead Cas9 protein (dCas9, grey) to genomic loci of interest, here regulatory elements such as gene promoters or distal enhancers (blue), in order to regulate expression of a specific gene (green).

In this way (and as I have described in [73]), it has been established that a dCas9-p300 histone acetyltransferase fusion can activate transcription of *MYOD* and *OCT4* both from proximal promoters and distal enhancers. At some of the sites tested, induction of mRNA following epigenome engineering with dCas9-p300 is stronger than activation achieved by a transcriptional activator at the same site [74]. Furthermore, it has been shown that demethylation of several (but not all) sites targeted in the *RHOXF2* promoter using a TALE-TET1 fusion leads to transcriptional up-regulation of this gene [75]. Addition of H3K4me3 by using the histone methyltransferase PRDM9 targeted via dCas9 or a zinc finger could in some cases achieve re-expression of silenced target genes [76]. Conversely, lysine demethylase LSD1 has been used to silence genes by targeting to known enhancer regions [77, 78]. Several targetable DNMT3a constructs have been reported and have been shown to decrease transcript levels when targeted to promoters [79–82]. Thus in summary, targetable chromatin modifiers have been used both to up- and to down-regulate mRNA levels. The experiments described above further provide direct evidence that chromatin modifiers can regulate transcription. In most cases, an effect on transcription could be detected following modification of some, but not all, targeted sites. This suggests inherent differences in the regulatory potential of genomic loci as well as at the level of individual chromatin marks. Consistent with results from genetic experiments, this implies that certain chromatin marks may only be functionally relevant at a subset of sites they occur at. One important unanswered question is whether the observed transcriptional changes are indeed mediated directly via changes in chromatin marks. Chro-

matin modifiers such as histone acetyltransferases (HATs) and histone deacetylases (HDACs) have been shown to have many non-histone substrates, including transcription factors such as p53, ETS and SMAD7 [83]. It is thus also possible that the observed effects could be relayed by local post-translational modification of transcription factors and this requires further investigation.

The functional consequences of epigenome engineering can be assessed by looking at changes in transcript level, protein level, or cellular phenotype. It is difficult to judge whether statistically significant but relatively small engineered changes in transcript level reported can be biologically relevant. However, if engineered marks translate into alterations of protein levels it is possible that they may indeed influence cellular behaviour. Several studies have already reported changes in protein level following epigenome engineering [78, 79, 84]. Ultimately, it will be important to test directly for changes in cellular or organismal phenotypes. A few studies have made such a connection already. It has been reported that addition or removal of single chromatin marks is sufficient to alter cell proliferation and colony-forming ability of cancer cells [79], the capacity for self-renewal of pluripotent stem cells [78] and even addiction-related behaviour in mice [84].

Important questions that remain are how common functional chromatin marks are and whether engineered changes to transcription can be maintained by cells and may even be heritable (see also [73]). While DNA methylation is generally thought of as a heritable and stable mark, there is emerging evidence that cells may counteract engineered changes. In one study, engineered DNA methylation marks were found to reduce to background levels *in vitro* [82] indicating they are either actively or passively lost. However, in another report targeting a different locus, engineered DNA methylation marks were found to persist [85]. Since the loci that were targeted differed in the two studies (and in the latter was located on a human artificial chromosome) it is possible that endogenous chromatin "context" determines whether an engineered change can be maintained. However, this still requires further investigation.

In line with the above argument, another recent study concluded that both the level of activation achieved by epigenome engineering as well as its stability of the engineered change through mitosis depends on chromatin environment [76]. The authors targeted four loci that they characterised as susceptible silenced, i.e. repressed without DNA methylation (*PLOD2* in C33a cancer cells, *EpCAM* in HEK293T and A549) or insusceptible, i.e. repressed and with DNA hypermethylation (*ICAM1* and *RASSF1a* in both HEK293T and A549) with the histone

methyltransferase PRDM9 in different cell lines. The dCas9 fusion protein did not appear to bind the hypermethylated sites efficiently as indicated by ChIP whereas the smaller zinc-finger fusions achieved up 60 % increase in H3K4me3 levels even at the promoters with DNA hypermethylation, indicating efficient binding. In both cases, however, increases of mRNA levels were modest. One inducible targetable PRDM9 construct achieved 8-fold increase in mRNA at the *EpCAM* promoter, however the effect was not sustained and expression decayed to background levels after 7 days in culture. Modification of the unmethylated *PLOD2* promoter with two different zinc-finger fusions appeared to be stable over the same time-course.

Another recent study reported remarkably strong and stable gene silencing of the *B2M* locus in K562 cells using a combination of targetable repressors [86]. The authors found that targeting the promoter of *B2M* endogenously tagged with tdTomato using a triple combination of DNMT3a, DNMT3L and KRAB (targeted either via dCas9 or a TALE DNA binding domain), achieved 500-fold reduction in mRNA levels and loss of Tomato expression in up to 78 % of cells. In HEK293T cells targeting the same locus only resulted in silencing in up to 25 % of cells, suggesting cell-type specific differences in the efficiency of epigenome editing. However, silencing using all three modifiers, but not individual constructs or combinations of two, was found to be stable over 50 days. Silencing was accompanied by complete methylation (100 %) at a number of CpG sites at the targeted locus, as well as loss of H3K4me3 and RNAPII signal with a concomitant increase of H3K9me3. Maintenance of silencing of this endogenous gene appeared to be dependent on DNA methylation. Expression of the gene could be reactivated by treatment with 5-azacytidine, a global inhibitor of DNA methyltransferases, or using targeted demethylation with a dCas9-TET1 construct while strong transcriptional activators such as dCas9-p300 or dCas9-VP160 or treatment with interferon-γ had no effect.

CM	Function	Full length or catalytic domain (CD)	targeted via	targeted to	Observed modification and effect size	Effect on transcription and effect size	Phenotypic effect (protein or other)	Locus tested	Model system/ Cell lines	Ref.
DNMT3a	DNA methyl-transferase	CD (amino acids 598-908)	ZFP	promoter	increased DNA methylation	60 % downregulation	protein (up to 80 %), reduced colony formation, reduced proliferation	*MASPIN*, *SOX2*	SUM159, MCF7	[79]
DNMT3a	DNA methyl-transferase	amino acids 602-912	dCas9	promoter	increased DNA methylation	40-50 % downregulation	NA	*IL6ST*, *BACH2*	HEK293	[82]
DNMT3a	DNA methyl-transferase	full length	dCas9	promoter, CTCF site	up to 70 % increase in methylation (reporter), 35 % at the *miR290* CTCF site, up to 40 % at *Pou5f1* CTCF site	silencing of GFP in up to 70 % of expressing cells, up to 3 fold increase in expression of some genes inside or outside the gene loop	loss of GFP (FACS)	Snrpn-GFP reporter inserted into *GAPDH* promoter, CTCF binding sites in *miR290* and *Pou5f1* gene loops	mESCs	[87]
DNMT3a-DNMT3L	DNA methyl-transferase	Dnmt3L C-terminal domain, Dnmt3a (CD)	ZFP, TALE	CpG island	increased DNA methylation	40-60 % downregulation	increased proliferation	*VEGF-A*, *p16*	SKOV3, HeLa, primary human fibroblasts	[88]
triple combination of targetable DNMT3a, DNMT3L and the KRAB repressor	DNA methyl-transferase	full length DNMT3L, Dnmt3a (CD)	dCas9, TALE	promoter, enhancer	up to 100 % DNA methylation, loss of H3K4me3 and RNAPII signal, increased H3K9me3 (*B2M*)	500-fold reduction in *B2M* mRNA, up to 80 % reduction in *IFNAR1* and *VEGFA*	stable loss of B2M-tdTomato only with triple combination of repressors in 78 % (K562) and 25 % (HEK293T) of cells determined by FACS, loss of MHC-I expression on cell surface	*B2M* (endogenously tagged with tdTomato), *IFNAR1*, *VEGFA*	K562, HEK293T	[86]

Table 1.1: Targetable chromatin modifiers currently available for epigenome engineering (gene repression) [73]

CM	Function	Full length or catalytic domain (CD)	targeted via	targeted to	Observed modification and effect size	Effect on transcription and effect size	Phenotypic effect (protein or other)	Locus tested	Model system/ Cell lines	Ref.
G9a	HMT	amino acids 829-1210	ZFP	promoter	increased H3K9 methylation (up to 2.7 fold)	40 % loss of mRNA	NA	*VEGF*	HEK293	[89]
HDAC8 (X.laevis), RPD3 (S.cerevisiae), Sir2a (P.falsiparum), Sin3a (H.sapiens)	histone deacetylase	amino acids 1-325 (HDAC8), amino acids 19- 340 (RPD3), amino acids 1-273 (Sir2a)	TALE	promoter	H4K8Ac reduction up to 50-60 %	up to 60 % decrease in RNA level	NA	*Grm2*	primary neurons	[90]
KYP (A. thaliana), SET8 (T.gondii), NUE (C. trachomatis)	histone methyltransferase	amino acids 1-331 (KYP), amino acids 1590-1893 (SET8), full length (NUE)	TALE	promoter	KYP: increased H3K9me1 (ca 1.4 fold), SET8: increased H4K20me3 (ca 2.4 fold) , NUE increased H3K27me3 (ca 2.2 fold)	50-75 % decrease in RNA level	NA	*Grm2*	primary neurons	[90]
LSD1	histone H3K4 demethylase	full length	TALE	enhancer	65 % loss of H3K4me2 and 60 % loss of H3K27ac (relative to TALE alone and scrambled TALE controls), up to 80 % loss in H3K4me2 and 90 % loss of H3K27ac measured relative to an mCherry transfection control	up to 50 % decrease in RNA level	NA	candidate enhancer in SCL (stem cell leukaemia) locus and 40 additional candidate enhancers, effect on transcription monitored for known targets or nearest expressed genes	K562	[77]
LSD1	histone H3K4 demethylase	full length	dCas9	enhancer	up to 85 % H3K4me2 loss, 90 % loss of H3K27ac	loss of mRNA	ES cell morphology changes	*OCT4* distal enhancer, 8 candidates enhancers thought to regulate pluripotency in ESCs, *Tbx3*	mESCs	[78]

Table 1.1: Targetable chromatin modifiers currently available for epigenome engineering (gene repression) [73] - continued

CM	Function	Full length or catalytic domain (CD)	targeted via	targeted to	Observed modification and effect size	Effect on transcription and effect size	Phenotypic effect (protein or other)	Locus tested	Model system/ Cell lines	Ref.
SID4X (4x mSin3 interaction domains)	H3K9 acetyl-transferase recruitment	NA	TALE (light-inducible)	promoter	50 % loss of H3K9ac	50 % loss reduction in mRNA	NA	*Grm2*	primary neurons	[90]
Sin3a (H.sapiens)	histone deacetylase recruitment	amino acids 524-851 (Sin3a)	TALE	promoter	H4K9Ac reduction by 30 %	75 % decrease in RNA level	NA	*Grm2*	primary neurons	[90]
SUV39H1	histone methyltrans-ferase	full length and shorter constructs	ZFP	Promoters	increased H3K9 methylation (up to 2.8 fold)	40 % loss of mRNA	NA	*VEGF*	HEK293	[89]

Table 1.1: Targetable chromatin modifiers currently available for epigenome engineering (gene repression) [73] - continued

CM	Function	Full length or catalytic domain (CD)	targeted via	targeted to	Observed modification and effect size	Effect on transcription and effect size	Phenotypic effect (protein or other)	Locus tested	Model system/ Cell lines	Ref.
p300	HAT	CD (amino acids 1,048 - 1,664)	dCas9, ZF, TALE	promoters and enhancers	increased H3K27ac (up to 10 fold increase relative to D1399 catalytic mutant	10 - 10,000 fold increase in RNA relative to D1399Y catalytic mutant	NA	promoters: *IL1RN*, *MYOD*, *OCT4*, β-globin (*HBE*, *HBG*), *ICAM1*, enhancers: *MYOD*, *OCT4*, β-globin	HEK293T	[74]
PRDM9	H3K4me3 HMT	CD ()	dCas9, ZF	targeted to	H3K4me3 increase (up to 60%)	upregulation of EpCAM (up to 8-fold in HEK293T) less successful activation at the other loci tested	NA	promoters of *ICAM1*, *RASSF1a*, *EpCAM*, *PLOD2*	HEK293T, A549, C33a	[76]
TET1	DNA demethy-lase	CD (amino acids 1,418-2,136)	TALE	promoters	10 - 80 % decrease in methylation levels (measured relative to off-target TALE or H1671Y, D1673A catalytic mutant	increase in transcription 2-1,000 fold relative to off target TALE or H1671Y, D1673A catalytic mutant	NA	*HBB*, *KLF4*, *RHOXF2*	K562, HEK293, HeLa	[75]
TET1	DNA demethy-lase	CD	dCas9	promoter (missing important controls, e.g. catalytic mutant)	up to 40 % demethylation	up to 2.5-fold increase in mRNA	slightly increased sensitivity to Mitomycin-C	*BRCA1*	HeLa, MCF7	[91]

Table 1.2: Targetable chromatin modifiers currently available for epigenome engineering (gene activation) [73]

CM	Function	Full length or catalytic domain (CD)	targeted via	targeted to	Observed modification and effect size	Effect on transcription and effect size	Phenotypic effect (protein or other)	Locus tested	Model system/ Cell lines	Ref.
TET1	DNA demethy-lase	CD	dCas9	promoter, enhancer	up to 60 % demethylation (reporter locus, *MYOD*), up to 35 % at *BDNF* promoter	3-fold increase in *BDNF-IV* and *MYOD* mRNA	re-expression of GFP reporter in 25 % of transduced cells (FACS) and in up to 70 % *in vivo* (lentiviral delivery to brain), expression of BDNF (fluorescence), fibroblast-to-myoblast conversion (*MYOD*, by immunofluorescence, required addition of 5-Aza)	Snrpn-GFP reporter inserted into *Dazl* locus, Snrpn-GFP inserted into the *Dlk1-Dio3* imprinted locus (paternally methylated, mouse model), *BDNF* promoter IV, *MYOD* distal enhancer	mESCs (reporter genes), post-mitotic neurons (*BDNF*), C3H10T1/2 MEF cells (*MYOD*), *in vivo* mouse model with paternally imprinted Snrpn-GFP reporter	[87]

Table 1.2: Targetable chromatin modifiers currently available for epigenome engineering (gene activation) [73] - continued

1.3 CRISPR-based epigenetic screening strategy

The aim of my PhD project is to develop an epigenetic screening method to identify genomic loci where addition or removal of a chromatin mark has a functional impact on cellular phenotype. To this end, I re-purposed the CRISPR system for use in an epigenetic screen (**Figure 1.6 B**, see also **Figure 1.8** for a graphical overview of the stages of the project).

This involves the construction of pooled libraries of gRNAs that each target a dCas9-chromatin modifier fusion protein to a specific site in the genome. In any given cell, co-expression of the dCas9-chromatin modifier and a single gRNA from the library will result in chromatin modification at a single site. If this modification has a significant effect on gene expression, either upregulating or downregulating expression of a gene, it should be possible to observe a phenotypic effect in that one cell. By using a pooled library of gRNAs it is possible to interrogate many genomic loci in a single experiment. Following epigenome modification, individual cells of interest showing a phenotypic change following epigenetic engineering have to be isolated from the population of cells. This is followed by sequencing of the gRNA protospacer that determines the corresponding genomic target site of modification.

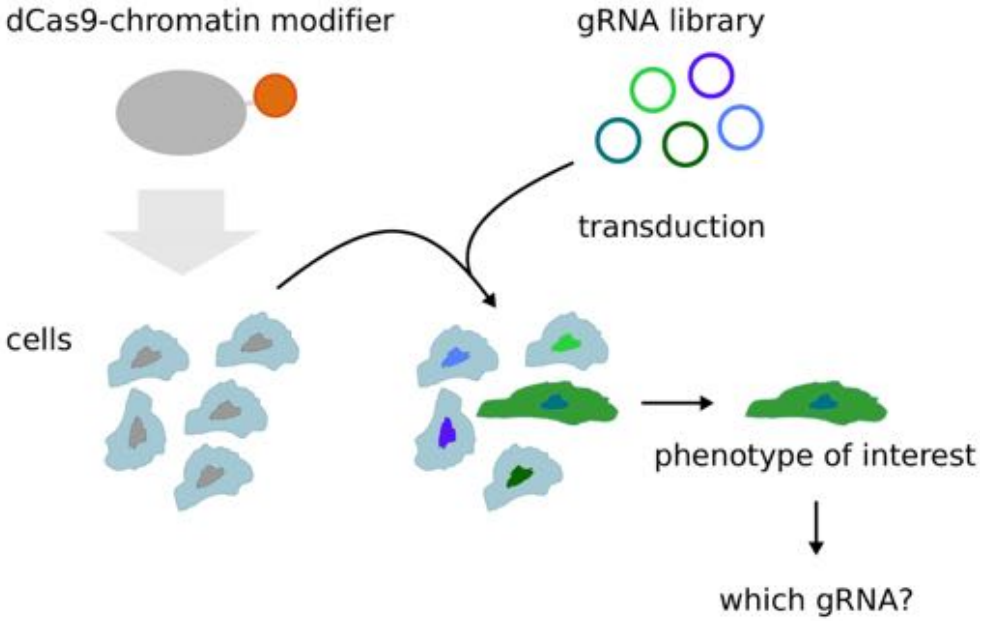


Figure 1.6: Screening strategy. A library of gRNAs is introduced into a cell line together with a dCas9-chromatin modifier construct. Following recovery leaving time for expression of the transgene, cells are analysed with respect to a particular phenotype. gRNAs are extracted from transfected cells. Comparing gRNAs present shortly after transfection and those present at the point of analysis, it is possible to make a statement about enrichment or depletion of gRNAs and therefore identify regions of interest that display activator/inhibitory action with respect to the chosen phenotype.

Once established, the CRISPR-based epigenetic screening method should be applicable to a plethora of phenotypes, provided an appropriate experimental read-out is available. An ideal phenotype to use as a read-out for an epigenetic screen should be clearly epigenetic, easy to assay using an established method and inducible *in vitro*. Furthermore, it is important to establish a method to isolate cells of interest, which may be based on antibiotic selection or fluorescence-activated cell sorting, for example. Perhaps the simplest application of such a screen would be to look for a gRNA/chromatin modifier combination that activates or represses expression of a reporter gene that is endogenously tagged with a fluorescent marker. A more complex phenotype to analyse would be migratory behaviour. For example, one could think of screening for activators or inhibitors of migration by using migration through filters as read-out.

1.3.1 CRISPR-based screening approaches to date

While a CRISPR-based screening method using chromatin-modifying enzymes has not been published to date, genetic CRISPR screens and screens using targetable transcriptional activators and inhibitors have already been reported [92]. These differ from an epigenetic screen in the design of the gRNA library. Genetic screens targeting protein-coding regions [93–98] use libraries based on oligonucleotide synthesis of gRNA protospacers designed *in silico*. 3-4 gRNAs per gene that preferentially target the first conserved exons and have a high predicted efficiency and specificity are included in the libraries. For the purpose of achieving a knock-out, every one of these 3-4 guides per gene included in the library is essentially equivalent - each one could in principle produce a knockout of the encoded protein when wild-type Cas9 induces a double strand break which is repaired by NHEJ, which in some cases lead to small insertions/deletions (indels) that produce a frameshift or nonsense mutation. When designing libraries for an epigenetic screen, adjacent sites may not be equivalent, i.e. a chromatin modifier might have to be targeted to a very specific site in, for example, a gene promoter to achieve the desired effect. This is more similar to the requirements for a library used in CRISPR-based activator/inhibitor screens (CRISPRa/i) [92]. For the activator screen, Gilbert *et al.* constructed genome-wide libraries comprised of roughly 200,000 gRNAs that target 15,977 human genes at 400 to 50 bp upstream from the transcriptional start site with up to 10 gRNAs per gene (TSS library, **Figure 3.19**). Cells stably expressing the sunCas9 system, a modified Cas9 construct that recruits multiple copies of the synthetic transcriptional activator VP64 to a single site, resulting in robust transcriptional activation [99] were used in the activator screens. For the inhibitor screen, gRNAs that fell into the region just downstream of the transcriptional start site (+50 to +100 bp) were included in the library. K562 cells expressing the dCas9 fused to the transcriptional repressor KRAB were used for the inhibitor screens. When combined, the two approaches - activator and inhibitor screens - allow screening for both loss-of-function and gain-of-function phenotypes. However, the libraries used in these approaches were largely focused around the TSS and are therefore of limited use when trying to assay larger regulatory regions inside and outside of gene promoters (**Figure 1.7**).

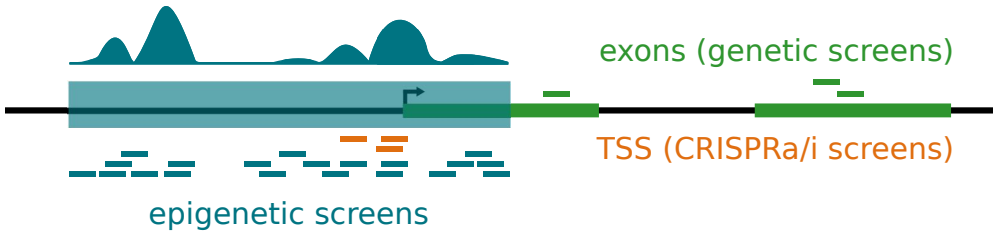


Figure 1.7: gRNA libraries suitable for CRISPR-based screens. CRISPR-based screens reported to date have used libraries containing gRNAs that target exons (green) or are focused around the transcriptional start site (TSS) of genes (orange). For an epigenetic screen, libraries that target larger regulatory regions where chromatin modifications have been identified, e.g. through profiling approaches, are required.

The gRNA libraries generated for this project differ in their complexity, targeting efficiency and coverage of the genome. The approaches taken explore the use of (1) a random gRNA targeting sequence, (2) degenerate consensus sequences that enrich for binding to promoters, (3) enzymatic digestion of genomic DNA down to roughly 20 bp fragments using an endonuclease and (4) oligonucleotide-synthesis of defined gRNA sequences designed *in silico* (see Methods and Results sections).

While saturating genome-wide genetic screens targeting all protein-coding genes in the human or mouse genome can now be conducted and these also show improved efficiency compared to RNAi/shRNA screens [97], saturating screens for non-coding regions have only been carried out for individual loci [100]. In one study, this approach was used to dissect the erythroid-specific intronic enhancer in the *BCL11A* gene to identify sites that, when mutated, lead to re-expression of foetal haemoglobin (HbF). A library containing all possible gRNAs along this enhancer was used and cells of interest were isolated using fluorescence-activated cell sorting (FACS) based on HbF expression level. Among the top hits from this screen was a gRNA that mapped directly onto a GATA1 motif whose loss renders the enhancer nonfunctional. Whether this method can be extended to screen for functional non-coding elements genome-wide remains to be determined.

One obvious challenge *en route* to genome-wide epigenetic screens will be to generate libraries that cover many more sites than were interrogated by the screens described above. To date, libraries typically contain up to 200,000 different gRNA sequences generated through oligonucleotide synthesis, which is relatively costly. Most epigenomic profiles contain many more peaks distributed across the entire genome than could be interrogated using this type of gRNA library. One part of my PhD thesis is to explore new methods for generating ultra-complex gRNA libraries.

One currently available option to identify sites in the genome where chromatin modification matters is pre-selection of loci of interest. In one study, several published profiles of chromatin marks (e.g. H3K4me2, H3K27ac) and transcription factor binding sites were integrated to generate a list of candidate active enhancers bound by TP53 [101]. To reveal which of these enhancers are necessary for one specific function of TP53, namely induction of oncogene-induced senescence, the authors introduced targeted mutations in 685 regions and found that most of the TP53 bound enhancers are dispensable for triggering senescence. Only two genomic binding-sites of TP53 were required for senescence induction.

1.3.2 Phenotypic readout

I identified cadherin switching as a potential phenotypic read-out for an epigenetic screen. Cadherin switching, specifically downregulation of E-Cadherin and concomitant upregulation of N-Cadherin, is a hallmark of epithelial-to-mesenchymal transition (EMT), a process thought to play a role in cancer metastasis [102, 103]. The process of metastasis involves detachment of a cancer cell from the primary tumour, migration through the blood stream and invasion and colonisation of a distal tissue. One of the first phenotypic changes to emerge is the disruption of the tight epithelial cell-to-cell contacts with concomitant loss of polarity and remodelling of the cytoskeleton. Changes in migratory behaviour need to be at least partially reversible in order to allow the cell to re-attach at the site of the secondary tumour [104, 105]. The molecular mechanisms underlying metastasis are complex and highly dynamic and - given the need for reversal of changes - cannot be brought about through mutation alone [106]. The metastatic program is also regulated by external and internal signalling that orchestrates complex behaviour through changes in gene expression.

Cancer cells are thought to undergo EMT when detaching from the primary tumour and extravasating to the blood stream and then have to undergo the reverse process, mesenchymal-to-epithelial transition (MET), when leaving the bloodstream to colonise a distal site. E-Cadherin and N-Cadherin expression can serve as a marker for EMT. Overexpression of exogenous N-cadherin increases motility, invasion and metastatic potential in human breast cancer cells [107]. Loss of expression of E-cadherin, which often coincides with the appearance of DNA methylation in the E-cadherin promoter, correlates with cancer progression in mouse models [108]. Loss of E-cadherin through mutation occurs only in some tumours which exert high invasive potential (subsets of breast and gastric can-

cers) [109]. Reinstatement of E-cadherin expression in cancer cell lines reduces cellular motility and invasiveness [110]. However, whether EMT underlies metastasis *in vivo* is still under debate [111]. EMT can be triggered *in vitro* through overexpression of various transcription factors, originally characterized in developmental systems. These include SNAIL, SLUG and TWIST [112]. Partial EMT has also been reported. For establishing the screen it is not important whether the cells undergo a true EMT. Downregulation of E-Cadherin and upregulation of N-Cadherin alone would be a sufficient phenotypic change for a positive read-out whether it is brought about by influencing expression of the two genes through EMT or another pathway. Cadherin switching can be induced in culture through addition of an EMT-inducing cell culture supplement containing TGF-β. Antibodies suitable for FACS are available against both E- and N-Cadherin, making it easy to assay this cellular phenotype and identify cells that have undergone a change in cadherin expression. The E-/N-Cadherin system thus appears to be a suitable read-out for establishing an epigenetic screening method.

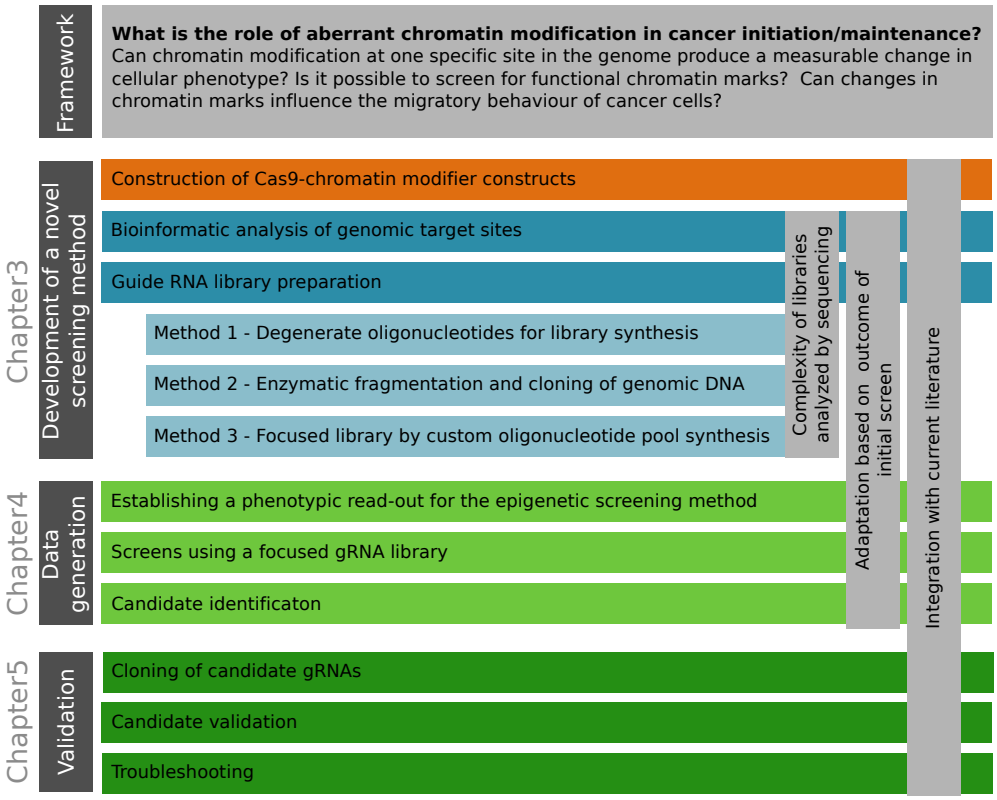


Figure 1.8: PhD Thesis outline and graphical overview of the different parts of the PhD project.

Chapter 2

Methods

2.1 dCas9-chromatin modifier fusion constructs

2.1.1 Construction of dCas9-chromatin modifier constructs

All Cas9 constructs are based on pMLM3705 (Addgene plasmid 47754). For addition of Hygromycin resistance cassette, pMLM3705 was cut with *SgrD*I (Thermo Scientific, ER2031) and *Mlu*I (Thermo Scientific, ER0561) and a Hygromycin resistance cassette inserted via Gibson cloning. The PGK promoter and Hygromycin resistance gene were cloned from Addgene Plasmid 41721 (MSCVhygro-F-G9a) followed by an SV40 early polyadenylation signal from the pCDNA3.1 backbone of Addgene Plasmid 13820 (HDAC1 Flag), primers used for PCR amplification can be found in **Table 2.1**. The resulting plasmid is named p-dCas9-VP64-Hygro.

Name	Sequence
Cas9-Hygro-cassette-1F	AGTCAATAATCAATGTCAACCGGGTAGGGGAGGCG
Cas9-Hygro-cassette-1R	GGTGGGCGAAGAACTCTCGGCATCTACTCTATTCCT TTG
Cas9-Hygro-cassette-pA-1F	GAATAGAGTAGATGCCGAGAGTTCTTCGCCCACCCC
Cas9-Hygro-cassette-pA-1R	AAGTGCCACCTGACGTCGACGGGTATACAGACATGA TAAGATACATTGATGA

Table 2.1: Primers for insertion of the Hygromycin cassette into pMLM3705

For exchange of VP64 with another domain (TET1, LSD1, DNMT3a, SET7/9, G9a, HDAC, JMJD2A, p300) the plasmid backbone was cut with *Pst*I (NEB) and *Pme*I (NEB) and fragments containing the relevant chromatin modifiers were inserted via Gibson cloning. Primer sequences used for PCR amplification are shown in **table 2.2** below. The human TET1 catalytic domain (a.a 1418-2136) was amplified from Addgene plasmid 39454 (pAAV-EF1a-HA-hTet1CD-WPRE-PolyA) and inserted in the MLM3705-Hygro backbone to yield p-dCas9-TET1-Hygro. For exchange of VP64 with the LSD1 domain the LSD1A isoform B catalytic domain (a.a. 171-852) was cloned from LSD1 cDNA clone IRATp970A0364D (Source Bioscience), yielding p-dCas9-LSD1-Hygro. The P405H change in this cDNA clone was reverted to the consensus during Gibson assembly.

Name	Sequence
TET1-CD-1F	CGATGACAAGGCTGCAGGAGGCGGAGGTAGCGAACTGCCC ACCTGCAGCTG
TET1-CD-int-1R	GGCAGTGACGAAGGCTTACT
TET1-CD-int-2F	AGTAAGCCTTCGTCACTGC
TET1-CD-2R	GCTGATCAGCGGGTTTTCAGACCCAATGGTTATAGG
LSD1-1F	CGATGACAAGGCTGCAGGAGGCGGAGGTAGCCCATCGGGT GTGGAGGG
LSD1-int-1R	AGGGACACAGGCTTATTATTGAGG
LSD1-int-1F	CCTCAATAATAAGCCTGTGTCCCT
LSD1-1R	GCTGATCAGCGGGTTTTCACATGCTTGGGGACTGCT
DNMT3a-CD-1F	CGATGACAAGGCTGCAGGAGGCGGAGGTAGCAACCACGAC CAGGAATTTGAC
DNMT3a-CD-1R	GCTGATCAGCGGGTTTTCAATACTCCTTCAGCGGAGCG
Set7-CD-1F	CGATGACAAGGCTGCAGGAGGCGGAGGTAGCTTCTTCTTT GATGGCAGCACC
Set7-CD-1R	GCTGATCAGCGGGTTTTCACTTTTGCTGGGTGGCC
G9a-1F	CGATGACAAGGCTGCAGGAGGCGGAGGTAGCGGCTATGAG AACGTGCC
G9a-1R	GCTGATCAGCGGGTTTTCATGTGTTGACAGGGGG
JMJD2A-1F	CGATGACAAGGCTGCAGGAGGCGGAGGTAGCGCTTCTGAG TCTGAAACTCTGAATCC
JMJD2A-1R	GCTGATCAGCGGGTTTTCATGCTTCTGGCGTGGGCAG
HDAC1-1F	CGATGACAAGGctgcaggaggcggaggtagcACGCAGGGCACCCGGA
HDAC1-1R	GCTGATCAGCGGGTTTTCAAATCGCCTGCATTTGGACCC
P300-1F	CGATGACAAGGCTGCAGGAGGCGGAGGTAGCAAAGAAAAT AAGTTTTCTGCTAAAAGG
P300-1R	GCTGATCAGCGGGTTTTCAGCATTCATTGCAGGTGTAGACA AA

Table 2.2: Primers for amplification of various chromatin modifier domains

For insertion of the DNMT3a domain, a region encompassing amino acids 598-908 was cloned from Addgene plasmid 36941 (pcDNA3/Myc-DNMT3A2) and

inserted, yielding p-dCas9-DNMT3a-Hygro. Set7/9 (a.a. 52 - 366) was amplified from Addgene plasmid 24082 (pET28 Set9 wt), G9a (a.a. 621-1000) was cloned from Addgene plasmid 41721 (MSCVhygro-F-G9a) and HDAC1 (a.a. 4-384) was cloned from Addgene plasmid 13820 (HDAC1 Flag). JMJD2A (a.a. 1-350) was amplified from Addgene plasmid 38846 (JMJD2A), p300 (a.a. 1284-1673) was amplified from Addgene plasmid 23252 (pcDNA3.1-p300). The resulting constructs were named p-dCas9-Set7-Hygro, p-dCas9-G9a-Hygro, p-dCas9-HDAC1-Hygro, p-dCas9-JMJD2A-Hygro and p-dCas9-p300-Hygro. All constructs were validated by restriction enzyme digest and Sanger sequencing.

2.1.2 Construction of dCas9-d-chromatin modifier mutant constructs

Constructs were made by mutagenesis PCR using the following primers together with the relevant Cas9-chromatin modifier constructs described above as template:

Name	Sequence
TET1-CDmut-H1672Y-D1674A-1F	GTGCTCATCCCTACAGGGCCATTCACAACAT
TET1-CDmut-H1672Y-D1674A-1R	ATGTTGTGAATGGCCCTGTAGGGATGAGCAC
TET1-CDmut-H1672Y-D1674A-2F	TCATCCCTACCGGGCCATTCACAACATGAATA ATGGAAGC
TET1-CDmut-H1672Y-D1674A-2R	TGTGAATGGCCCGGTAGGGATGAGCACAGAAG TCCAG
DNMT3a2-CD-C706S-mut-1F	GTCCTTCGAATGACCTCTCCATCGTCAACCCT GCT
DNMT3a2-CD-C706S-mut-1R	GGTCATTCGAAGGACTGCCCCCAATCACCAGA
Set7-CD-H297G-mut-1F	GCAAATGGATCCTTCACTCCAAACTGCATCTA CGATATGTTTGTCCAC
Set7-CD-H297G-mut-1R	GAGTGAAGGATCCATTTGCCTTGTGTCCCAAG GAGGCACAG
LSD1-CD-K661A-mut-1F	CAACCTTAACGCAGTGGTGTTGTGTTTTGATC GGGTG
LSD1-CD-K661A-mut-1R	ACAACACCACTGCGTTAAGGTTGCCAAATCCC ATCC
G9a-CD-Y1154A-1F	GTTTGACGCAGGCGATCGATTCTGGGACATCA AAAGCAAATATTTCAC

G9a-CD-Y1154A-1R	GTCCCAGAATCGATCGCCTGCGTCAAACCCTA GCTCCTCC
JMJD2A-CD-S288G-T289G-1F	CTGTGCGGAGGGTGGCAATTTTGCTACCCGTC GGTGGATTGA
JMJD2A-CD-S288G-T289G-1R	ACGGGTAGCAAAATTGCCACCCTCCGCACAGT TAAAACCATGGTTAAAG
p300-CD-S1396R-S1397R-1F	GAGAGTATATATAAGAAGGCTCGATAGTGTTC ATTTCTTCC
p300-CD-S1396R-S1397R-1R	CACTATCGAGCCTTCTTATATATACTCTCCTCT GGTT
HDAC1-CD-H141A-1F	GGGCCTGCATGCTGCAAAGAAGTCCGAGGCAT CTGGC
HDAC1-CD-H141A-1R	CTTCTTTGCAGCATGCAGGCCCCCAGCCCAAT TCACAGC

Table 2.3: Primers for making chromatin modifier mutants

Reactions were set up as follows: 25 μl 2X Phusion High Fidelity Master Mix (NEB, M0531S), 0.5 μM of each primer, 80 ng plasmid DNA as template and 1.5 μl DMSO. A touchdown reaction was performed using the following cycling conditions: 1 cycle at 98 °C for 30 s, 98 °C for 10 s, 98 °C (-1 °C per cycle down to 63 °C) for 10 s, 72 °C for 3 min, then 10 cycles of 98 °C for 10 s, 72 °C for 3 min, and final elongation at 72 °C for 10 min. Reactions were purified using Agencourt AMPure XP beads (Beckman Coulter, A63881) and the plasmid template digested with 20 U *Dpn*I (NEB, R0176S), followed by another bead clean-up. PCR products self-assemble to form a nicked circle and were directly transformed into OneShot Top10 electro-competent E.*coli* (Invitrogen, C4040-10) according to the manufacturer's instructions. The constructs were validated by restriction enzyme digest and Sanger sequencing.

2.1.3 Western blotting and nuclear fractionation

HEK293T cells (ATCC 293T/17, CRL-11268) were grown in DMEM (Life Technologies, 41966052) supplemented with 10 % FBS (Sigma). Cells were transfected with dCas9-chromatin modifier plasmids using Lipofectamine LTX (Life Technologies). Three days after transfection cells were harvested and lysed in CelLytic

M (Sigma, C2978) lysis buffer for whole-cell extracts. For nuclear and cytoplasmic fractionation, the NE-PER Nuclear and Cytoplasmic Extraction Reagents (Thermo Scientific, 78833) was used according to the manufacturer's protocol.

For analysis of samples by Western blot, 2X Laemlli buffer was added to samples, followed by incubation at 95 °C for 10 min and SDS-PAGE on an 8 % gel (run at 85V). Proteins were transferred to a PVDF membrane at 35 V for 16h at 4 °C/or 65V for 2hours. To facilitate transfer of large proteins, 0.05 % SDS was added to the transfer buffer. Membranes were blocked in TBST with 5 % milk protein (Sigma) for several hours at 4 °C, incubated with primary antibody at 4 °C overnight, washed 3x with TBST (10 min each) and incubated with mouse secondary antibody (1:5,000) for 1h at room temperature followed by 3 washes with TBST. Staining was visualised by incubation with home-made Detection Reagent (equal volumes of ECL solution 1 (2.5 mM Luminol, 0.4 mM pCoumaric acid, 100mM Tris pH 8.5) and ECL solution 2 (0.01 % H_2O_2, 100mM Tris pH 8.5)) for 2 min followed by exposure to a High Performance Chemiluminescence Film (GE Healthcare).

Antibody concentrations: anti-Flag M2 (Sigma) at a dilution of 1:1,000, anti-MEK1 H-8 (sc-6250) at a dilution of 1:500, anti-p53 (1C12) at a dilution of 1:1,000 anti-Tubulin (YL1/2) at a dilution of 1:1,000

2.1.4 *F*unctional validation of dCas9-TET1

HeLa cells were maintained in DMEM (Life Technologies, 41966052) supplemented with 10 % FBS (Sigma). 6x10^5 HeLa cells were seeded in a well of 6-well plates and transfected with (1) 100 ng GFP, (2) 300 ng dCas9-TET1 (wild-type) and 300 ng gRNA or (3) 1.2 μg RH3 TALE-TET1 (pMLM3709, Addgene plasmid 49943) the next day. The gRNA sequences for the RH3-1 gRNA construct is: CTGTGGGTTGGGCCTGCTG. 3 μl Lipofectamine LTX (Thermo Scientific, 15338100) were used per transfection reaction in all cases except for RH3 TALE-TET1 construct, where 3.3 μl Lipofectamine LTX and 1 μl Plus reagent was used according to ref. [75].

On day 3 after transfection, cells were harvested. 75,000 cells per condition were used for DNA extraction and bisulfite-conversion using the EZ DNA Methylation-Direct Kit (Zymo Research, D5020) according to the manufacturer's instructions. The target region was amplified from bisulfite-converted DNA using primers RhoxF2-bis-1F (TGGGGTTAGTGTAAAATTTTTTTTAT) and RhoxF2-bis-

1R (CAATATATCCACTTAAAAACCTCCTCT) and the PyroMark PCR Kit (Qiagen, 978703). 12.5 μl 2X PyroMark Master Mix, 2.5 μl 10X Coral Load, 0.2 μM of each primer and 15 ng bisulfite-converted DNA were used in the reaction. Cycling conditions were as follows: 1 cycle at 95 °C for 15 min, 45 cycles of 94 °C for 30 s, 60 °C for 30 s, 72 °C for 30 s and final elongation at 72 °C for 10 min and A-tailing at 60 °C for 30 min. Reaction were bead-purified using Agencourt AMPure XP (Beckman Coulter, A63881) according to the manufacturer's instructions. Fragments were cloned into pCR2.1 at a vector:insert ratio of 1:3 using the TA cloning kit (Invitrogen, K204040) according to the manufacturer's instructions. Individual colonies were picked for colony PCR the next day. Reactions were set up as follows: 25 μl 2X Long Amp Taq Master Mix (NEB, M0287S), 2 μl of 10 μM primer M13-F20 (GTAAAACGACGGCCAGTG), 2 μl of 10 μM primer M13-26Rev (CAGGAAACAGCTATGAC), and water up to 50 μl. Cycling conditions were: 1 cycle at 95 °C for 5 min, 40 cycles of 95 °C for 15 s, 59 °C for 15 s, 65 °C for 50 s and final elongation at 65 °C for 5 min. 15 μl of the reaction were analysed on a 1.5 % gel stained with Sybr Safe stain (Invitrogen). Reactions yielding amplicons of the correct size were treated with ExoSAP-IT (Affymetrix, 78200, 4.4 μl ExoSAP-IT per 11 μl PCR product) and sequenced by Sanger sequencing.

2.1.5 *F*unctional validation of dCas9-chromatin modifier constructs other than dCas9-TET1: Immunoprecipitation and *in vitro* activity assays

HEK293T cells (ATCC 293T/17, CRL-11268) were grown in DMEM (Life Technologies, 41966052) supplemented with 10 % FBS (Sigma). 2x10^6 cells were seeded per well in 6-well plates and transfected with Lipofectamine LTX and 7 μg plasmid DNA. Cells were selected with 100 μg/ml Hygromycin. Immunoprecipitation of dCas9-chromatin modifier constructs was performed using the FLAG Immunoprecipitation kit (Sigma). 1x10^7 or 2x10^7 cells (for the 2X lysate sample) were trypsinised, washed twice with PBS and lysed in 1 ml Lysis Buffer (50 mM Tris HCl, pH 7.4, with 150 mM NaCl, 1 mM EDTA, and 1 % TRITON X-100) for 30 min. All remaining steps were carried out at 4 °C. 20 μl of ANTI-FLAG M2 affinity gel was used per immunoprecipitation reaction. The resin was prepared by washing twice with 0.5 ml Wash Buffer (50 mM Tris HCl, pH 7.4, with 150 mM NaCl). An optional wash with Elution buffer (0.1 M Glycine, pH 3.5) to remove unbound anti-Flag antibody was also included and the gel resin washed an addi-

tional three times with Wash Buffer. The entire lysate was added to the resin. For the FLAG-BAP positive control, 250 ng FLAG-BAP protein was added to 1 ml wash buffer and incubated with the resin. Samples were agitated for 2 hours at 4 °C, followed by 3 washes with Wash Buffer. Different elution methods were tested including elution with 2x 50 μl FLAG-elution buffer (150 ng/ μl 3X FLAG peptide in Wash Buffer), for 30 minutes at 4 °C. Using 100 μl Elution Buffer (0.1 M Glycine, pH 3.5), elution was performed for 5 min at room temperature. For elution with Sample Buffer, 20 μl of 2X Sample Buffer (125 mM Tris HCl, pH 6.8, with 4 % SDS, 20 % (v/v) glycerol, 0.004 % bromophenol blue) was added to the pelleted resin and incubated at 100 °C for 3 min. Immunoprecipitates and supernatants were analysed by Western blotting using an the FLAG-M2 antibody (see section 2.1.3)

In vitro activity assays were performed using the following kits according to the manufacturer's instructions: EpiQuik DNMT Activity/Inhibition Assay Ultra Kit - Colorimetric (Epigentek Cat. P-3009), Epigenase 5mC Hydroxylase TET Activity/Inhibition Assay Kit - Colorimetric (Epigentek Cat. P-3086), EpiQuik Histone Methyltransferase Activity/Inhibition Assay Kit - H3-K9 (Epigentek Cat. P-3003), Epigenase JMJD2 Demethylase Activity/Inhibition Assay Kit - Fluorometric (Epigentek Cat. P-3081), EpiQuik Histone Methyltransferase Activity/Inhibition Assay Kit - H3-K4 (Epigentek Cat. P-3002), Histone Demethylase Assay - Fluorescent (version A1) (Active motif, Cat. 53200), HAT Assay Kit - Fluorescent (version B1) (Active motif, Cat. 56100), HDAC Assay Kit - Colorimetric (version B1) (Active motif, Cat. 56210)

2.1.6 Construction of lentiviral constructs

The plasmid plenti-dCAS-VP64-Blast was obtained from Addgene (Addgene Plasmid 61425) and digested with *BsiW*I (NEB) and *BsrG*I (NEB). The dCas9-VP64 variant from the previously constructed non-lenti plasmid p-dCas9-VP64-Hygro was amplified using primers dCas9-lenti-T2A-puro-1F and dCas9-VP64-lenti-T2A-puro-1R and inserted into the lenti backbone via Gibson cloning, yielding plasmid p-my-lenti-dCas9-VP64-T2A-Blast.

For the construction of p-lenti-dCas9-TET1-T2A-Blast, the plasmid p-my-lenti-dCas9-VP64-T2A-Blast was digested with *BamH*I-HF (NEB) and *Pac*I (NEB). A fragment containing TET1 was amplified from the previously made p-dCas9-TET1-Hygro using primers Cas9-univ-lenti-2F and Cas9-TET1-lenti-1R and inserted into the backbone via Gibson cloning.

For the construction of p-lenti-dCas9-p300-T2A-Blast, a PCR amplicon generated using primers Cas9-univ-lenti-1F and Cas9-p300-lenti-1R was inserted into p-lenti-my-dCas9-VP64-T2A-Blast that had been digested with *BamH*I (NEB) and *Pac*I (NEB).

p-lenti-dCas9-LSD1-T2A-Blast was generated by digesting p-my-lenti-dCas9-VP64-T2A-Blast with *BamH*I and *Pac*I and doing a 4-fragment Gibson assembly inserting a 2.1 kb fragment of p-dCas9-LSD1-Hygro produced during digestion of p-dCas9-LSD1-Hygro with *Xho*I and /*Dra*III-HF (NEB) together with two "sealing fragments" derived by pre-annealing oligonucleotides Cas9-BamH1-seal-1F and Cas9-BamH1-seal-1R, and dCas9-LSD1-lenti-seal-1F and dCas9-LSD1-lenti-seal-1R.

For the construction of p-lenti-dCas9-HDAC1-T2A-Blast, the plasmid p-my-lenti-dCas9-VP64-T2A-Blast was digested with *BamH*I-HF and *Pac*I. A fragment containing HDAC was amplified from the previously made p-dCas9-HDAC1-Hygro using primers Cas9-univ-lenti-1F and dCas9-HDAC1-lenti-1R and inserted into the backbone via Gibson cloning. For the construction of p-lenti-dCas9-DNMT3a-T2A-Blast, a fragment was amplified from plasmid p-dCas9-DNMT3a-Hygro using primers Cas9-univ-lenti-1F and dCas9-DNMT3a-lenti-1R and inserted into the backbone. For the construction of p-lenti-dCas9-SET7-T2A-Blast, a fragment was amplified from plasmid p-dCas9-SET7-Hygro using primers Cas9-univ-lenti-1F and dCas9-Set7-lenti-2R and inserted into the backbone. Primers Cas9-univ-lenti-1F and dCas9-G9a-lenti-1R were used to amplify G9a from p-dCas9-G9a-Hygro and the fragment was inserted into the p-my-lenti-dCas9-VP64-T2A-Blast backbone digested with *BamH*I-HF and *Pac*. For construction of p-dCas9-JMJD2A-T2A-Blast, JMJD2A was amplified from p-dCas9-JMJD2A-Hygro using primers Cas9-univ-lenti-1F and dCas9-JMJD2a-lenti-1R and inserted into p-my-lenti-dCas9-VP64-T2A-Blast. To construct a p-lenti dCas9-eGFP-Blast plasmid, p-my-lenti-dCAS-VP64-T2A-Blast was digested with AscI (NEB) and PacI (NEB). eGFP was amplified from p-eGFP-C1 (Invitrogen) and inserted into the cut backbone.

Name	Sequence
dCas9-lenti-T2A-puro-1F	ATTTCAGGTGTCGTGACGTACGGCCACCATGGATAAAAAGTATTCTATTGGTTTAG
dCas9-VP64-lenti-T2A-puro-1R	GCCCTCTCCACTGCCTGTACAGTTAATTAACATATCGAGATCGAAATCG
Cas9-univ-lenti-1F	GTCACAGCTTGGGGGTGA
Cas9-univ-lenti-2F	GTCACAGCTTGGGGGTG
Cas9-TET1-lenti-1R	CTGCCTGTACAGTTAATGACCCAATGGTTATAGG
Cas9-p300-lenti-1R	CTGCCTGTACAGTTAATGCATTCATTGCAGGTGTAG
Cas9-BamH1-seal-1F	AGCTTGGGGGTGACGGATCCCCCAAGAAGAAGAGG
Cas9-BamH1-seal-1R	CCTCTTCTTCTTGGGGGATCCGTCACCCCCAAGCT
dCas9-LSD1-lenti-seal-1F	GCCAGGCCACACCAGGTGTTCCTGCACAGCAGTCCCCAAGCATGATTAACTGTACAGGCAG
dCas9-LSD1-lenti-seal-1R	CTGCCTGTACAGTTAATCATGCTTGGGGACTGCTGTGCAGGAACACCTGGTGTGGCCTGGC
dCas9-HDAC1-lenti-1R	CTGCCTGTACAGTTAATAATCGCCTGCATTTGGAC
dCas9-DNMT3a-lenti-1R	CTGCCTGTACAGTTAATATACTCCTTCAGCGGAGCGAAGAG
dCas9-Set7-lenti-2R	CTGCCTGTACAGTTAATCTTTTGCTGGGTGGCCT
dCas9-G9a-lenti-1R	CTGCCTGTACAGTTAATTGTGTTGACAGGGGGC
dCas9-JMJD2a-lenti-1R	CTGCCTGTACAGTTAATTGCTTCTGGCGTGGGCAG
dCas9-eGFP-T2A-Blast-1F	GGCGGTGGAAGCGGGGTGAGCAAGGGCGAGG
dCas9-eGFP-T2A-Blast-1R	GCCTGTACAGTTAATCTTGTACAGCTCGTCCATGC

Table 2.4: PCR primers for Gibson cloning of dCas9-chromatin modifier fusions into a lentiviral backbone

2.1.7 Packaging of lenti-dCas9 constructs into lentivirus

Lentiviral packaging of dCas9 constructs and virus concentration was performed according to standard procedures by Catherine King, who provides this service to the Cancer Genome Engineering Facility (CAGE).

Briefly, HEK293T (ATCC 293T/17, CRL-11268) cells were cultured in complete DMEM (Life Technologies, 41966052) with 10 % FBS and transiently transfected with the lentiviral expression construct and the two packaging vectors p8.91 (gag-pol expressor) and pMDG (VSV-G expressor) using Fugene transfection reagent (Promega, E2691). Medium was changed the next day. Lentivirus-containing

medium was harvested 24 and 48 hours after this initial media change. The harvested medium was filtered through a 0.22 μm filter.

Virus-containing medium was further concentrated by ultracentrifugation (64,900rcf) through a 25 % sucrose cushion for 2 hours at 4 °C and subsequently stored in aliquots at -80 °C until further use.

2.1.8 Determination of viable virus titer in A549 cells

A549 cell lines were maintained in Ham's F-12K (Kaighn's) Medium (Gibco) with 10 % FBS and 1 % Pen/Strep (Gibco). 0.5x10^5 cells were seeded per well in 12-well plates the day before transfection. The next day, cells were transduced with virus at various concentrations in complete media containing with or without 10 μg/ml Polybrene (Millipore). The next day, cells were washed and trypsinised and each well split into two wells containing equal amounts of cells. Cells in one of the replica wells were plated in medium containing 1.5 μg/ml Puromycin (Gibco). On day 3 after transduction, cells in each of the wells were trypsinised and counted. Only wells containing <20 % of viable cells were used for the calculation of viable titer according to the following formula:

Viable titer in transducing units/ml (TU/ml) = % antibiotic positive cells x number of cells transduced/vol of virus added to well (ml)

were % cells transduced = number of cells in well + puromycin/ number of cells in corresponding well - Puromycin

2.1.9 Polyclonal stable cell lines

A549 cells were transduced with lentiviral particles containing the dCas9-chromatin modifier-T2A-Blasticidin expression cassette at an MOI of 1. The next day, selection with 30 μg/ml Blasticidin was started.

2.1.10 Monoclonal stable cell lines

Transduced A549 cells were plated into 96-well plates at single cell dilution (0.7 cells per well). Cells were maintained under selection with Blasticidin (30 μg/ml). Once colonies had grown from single cells, empty wells and wells with more than one colony were excluded from further analysis. Immunofluorescence staining was

used to identify the clonal lines with the highest expression levels of the relevant dCas9-chromatin modifier construct. These clones were selected for further study.

2.1.11 Immunofluorescence staining for expression of lentiviral dCas9-chromatin modifier constructs

Transduced A549 cells or control cells were seeded into a well of a 12-well plate the day before fixation. The next day, cells were washed 2x with PBS (Gibco, 5 min per wash) and fixed in 4 % paraformaldehyde in PBS for 10 min at room temperature. Cells were washed 2x with PBS (5 min per wash) and permeabilised with 0.25 % Triton-X-100 in PBS at room temperature for 5 min. Cell were washed 2x with PBS (5 min per wash) and incubated in blocking solution (10 % BSA in PBS) at room temperature for 1 hour. Next cells were incubated in 1° antibody (anti-FLAG-FITC, M2, Sigma) diluted 1:500 in 3 % BSA/PBS at 4 °C overnight. Wells were washed 3x with PBST (PBS with 0.1 % Triton-X-100, 5 min per wash) and stained with Hoechst 33342 (BD) at a concentration of 1 μg/ml in PBS for 2 min followed by one wash with PBST before imaging.

2.2 Construction of guide RNA plasmids

The plasmid pMLM3636 (plasmid ID 43860) was obtained from Addgene and modified to be suitable for use with rolling circle amplification. The vector was cut with *BsmB*I and a double-stranded DNA fragment, generated by annealing the two oligos MLM3636-1F (5'-ATCTTGTGGAAAGGACGAAACACCGGT TTTAGAGCTAGAAATAGCAAGTT) and MLM3636-1R (5'-AACTTGCTA TTTCTAGCTCTAAAACCGGTGTTTCGTCCTTTCCACAAGAT), inserted via Gibson cloning. This yields a modified vector, referred to here as pgRNA-Neo, that contains the U6 promoter, followed by the 5'G of the gRNA sequence, followed by the scaffold sequence.

For lentiviral vector construction, the vector pLKO.1 (Addgene plasmid 10878) was modified to insert a gRNA promoter and scaffolding sequence. The vector was first digested with *Eco*RI (NEB) and *Age*I (NEB). Next, the desired sequences was amplified from pgRNA-Neo using primers gRNA-PLKO-F (5'-TTTCTTG GGTAGTTTGCAGTTTT) and gRNA-PLKO-R (5'-CCATTTGTCTCGAGG TCGAGTACCTCGAGCGGCCCAAGC). The resulting vector is referred to as pgRNA-pLKO.1

2.2.1 Bioinformatic analysis of target regions and design of the degenerate gRNA libraries

A detailed documentation including the code used for this analysis can be found on github at `https://github.com/annakoe/AnalysisScripts`. Briefly, all possible gRNA sequences in the human genome that harbour an NGG PAM were found by searching for the nucleotide sequence "$GN_{20}GG$" using the Bioconductor package BSgenome [113]. These regions were intersected with a file of putative promoter regions. The annotation file of promoter regions was generated by downloading a list of coordinates of all known transcripts (human GRCh37-70) from Ensembl (version70) and parsed through a script which extracts coordinates -1000 and +500 bp from the start of the transcript (`https://github.com/regmgw1/regmgw1_scripts/blob/master/ensembl_scripts/transcript2promoter.pl`). From this file, non-overlapping promoter sequences were derived by strand-specific merging using the Bedtools suite (v2.17.0) [114]. Regions where gRNAs fall into promoters were defined as regions of interest (ROIs). ROIs were clustered based on sequence identity using LCS-HIT (Version 0.5.2)[115]. Clusters were ranked in descending order by the number of members and the top 15 clusters extracted. For each of the top 15 clusters consensus sequences were computed using the Bioconductor package Biostrings [116]. The threshold option allows the user to define the percentage threshold at which a given nucleotide will be incorporated into the consensus sequence at a given position. The threshold for choosing the consensus was varied empirically so as to yield a sequence complexity of between 10^4 and 10^6 different sequences for each of the cluster consensus sequences. The targeting efficiency (number of target hits divided by total number of sequences represented by the consensus) was calculated. Cluster sequences were ranked by targeting efficiency and the top 6 clusters chosen for synthesis.

2.2.2 Construction of degenerate libraries from cluster sequences

The following degenerate sequences of the form 5'-[Phos]-N_{19}-GTTTTAGAGCT AGAAATAGCAAGTTAAAATAAGG, where N_{19} denotes the degenerate cluster targeting sequence derived as described above, were ordered from Sigma:

The gRNA library were generated by using one of these oligos as forward primer

Name	Sequence
M2-Cluster190-T20	5'-[Phos]-RRRGRGRRRRRGRRGRRGVGTTTTAGAGCTAGAA ATAGCAAGTTAAAATAAGG
M2-Cluster369-T21	5'-[Phos]-GRRRGRGRRRRRGRRRSVRGTTTTAGAGCTAGAA ATAGCAAGTTAAAATAAGG
M2-Cluster422-T19	5'-[Phos]-RRRRRRRRRRRRRRRRRGRGTTTTAGAGCTAGAA ATAGCAAGTTAAAATAAGG
M2-Cluster304-T22	5'-[Phos]-GGKKKGKGKKGKKKGKKBSGTTTTAGAGCTAGA AATAGCAAGTTAAAATAAGG
M2-Cluster89-T22	5'-[Phos]-VRVSSSSSGSSGSSRGSSKGTTTTAGAGCTAGAAATA GCAAGTTAAAATAAGG
M2-Cluster1281-T25	5'-[Phos]-GRGRRRRRGRRRRRGSRRVGTTTTAGAGCTAGAA ATAGCAAGTTAAAATAAGG

Table 2.5: PCR primers for making the degenerate gRNA library using circle amplification

together with MLM3636-1R (5'-[Phos]-CGGTGTTTCGTCCTTTCCAC) in a circle amplification from plasmid gRNA-pLKO.1 in a 50 μl reaction with Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB), 2ng template and 2μl 100 μM primer. PCR program: 1 cycle 98 °C for 30s, 30 cycles 98 °C for 10s, 62 °C for 10s, 72 °C for 3.5 minutes, followed by cooling to 12 °C and addition of 1 μl T4 DNA polymerase (3 U/μl, NEB) and incubation for 20 min at 12 °C. The PCR products were purified using bead purification with Agencourt AMPure XP (Beckman Coulter, A63881) according to the manufacturer's instructions, proceeding immediately to ligation under dilute conditions (10 μl 10X ligase buffer (NEB), 1.3 μl concentrated T4 DNA ligase (2000 U/μl)) incubating at 16 °C for 16 h. Ligation was followed by digestions of template DNA with *Dpn*I. 11.4 μl CutSmart Buffer (NEB) and 4 μl *Dpn*I (NEB) were added to the ligation followed by incubation at 37 °C for 2hrs followed by purification with the MinElute Reaction Cleanup Kit (QIAGEN) and electroporation of the entire ligation reaction into freshly prepared electro-competent TG1 *E. coli* cells (original stock from Lucigen) with a competency $> 10^{10}$ colony-forming units per μg DNA as determined by control electroporation with pUC19 plasmid (NEB). *E. coli* cells were allowed to recover in antibiotic-free medium for 1h at 37 °C before plating on selective 2TY-coated plates (Bio-assay dish with lid, 245mm x 245mm x 25mm, radiation sterilized, Thermo Scientific Nunc). Following overnight incubation at 37 °C, the bacteria were harvested by scraping and the plasmid library extracted using the HiSpeed Plasmid Maxi Kit (QIAGEN).

2.2.3 Construction of gRNA libraries from MNase-digested genomic DNA

Human genomic DNA extracted from pooled blood (250 ng/μl, generous gift from Lee Butcher) or mouse genomic DNA (Promega) was digested with various amounts of micrococcal nuclease (NEB) to determine the optimal amount of enzyme that digests DNA to between 5 bp and 100 bp in size. The reaction setup was as follows: 1 μg genomic DNA, 1 μl 10X MNase Buffer, 0.1 μl 100X BSA in a 10 μl reaction volume was incubated with enzyme for 15 min at 37 °C. The enzyme was immediately inactivated through addition of 1 μl EGTA (500mM). Following addition of 4 μl gel loading dye (Invitrogen), the reactions were run on a 20% PAGE gel (Invitrogen). A band ranging from 15 to 30 bp was excised from the gel and the DNA extracted using the Crush-and-Soak method [117]. Briefly, the gel was crushed using a sterile pipette tip and incubated in PAGE solubilisation buffer (0.5 M ammonium acetate, 10 mM magnesium acetate, 1mM EDTA pH8) at 37 °C for 16h. DNA was then extracted from the buffer using standard Phenol Chloroform extraction [118] and the ends end-repaired with the Quick Blunting kit (NEB) in a 15 μl reaction according to the manufacturer's instructions. This was followed by phenol-chloroform extraction and EtOH precipitation.

Linkers for cloning the end-repaired DNA fragment into the vector pgRNA-pLKO.1 via a Gibson reaction were amplified from the vector pgRNA-pLKO.1 using primers 5'-linker-F (5'-TTGGAATCACACGACCTGGA) and 5'-linker-R (5'-CGGTGTTTCGTCCTTTCCAC) and 3'-linker-F (5'-GTTTTAGAGCTAG AAATAGCAAGTTAAAATA) and 3'-linker-R (5'-ACTCGGTCATGGTAAGC TCC) respectively. Reactions set-up was as follows: 25 μl Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB), 2.5 μl of each primer (100 μM), 0.1 ng pgRNA-pLKO.1 in a total reaction volume of 50 μl. Cycling conditions: 1 cycle 98 °C for 30s, 32 cycles 98 °C for 10s, 59 °C for 10s, 72 °C for 30 seconds, followed by final elongation at 72 °C for 10 min. Fragments were purified using Agencourt AMPure XP beads (Beckman Coulter, A63881).

The 5' linker (689 bp) was digested with *Hind*III and a 600 bp fragment purified from a 1 % agarose gel using the Gel Extraction kit (QIAGEN). The 3' linker (848 bp) was digested with *Sac*II and the resulting 300 bp fragment gel-purified. 14 μl ligation reactions were set up with either equimolar amounts of MNase-digested fragments (5 ng) to linkers using 1.4 μl concentrated T4 DNA ligase (NEB) and incubated at 16 °C for 16h. Next, and without heat-inactivating the enzyme,

ligation reactions were directly used in nick translation, supplementing with 25 μl Long Amp Taq 2X Master Mix (NEB) and 2.5 μl primer Linker-Minus450-F (10 μM, 5'-GGGCAAGTTTGTGGAATTGG) 2.5 μl primer Linker-Plus275-R (10 μM, 5'-AAGTGGATCTCTGCTGTCCC) in a 50 μl reaction. Cycling conditions were 1 cycle at 72 °C for 20 min, and 3 cycles of 95 °C for 5 min, 95 °C for 15s, 58 °C for 15s, 72 °C for 30s and final elongation at 72 °C for 5 min. Reactions were cleaned up with Agencourt AMPure XP (Beckman Coulter, A63881) using a sample:bead ratio of 1:1.

Next, three different-size fragments were amplified from the product of nick-translation, using Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB) in a 25 μl reaction supplemented with 2.5 μl of 100μM primer and $1/16^{th}$ of the purified product of nick-translation as input. Using primers Linker-Minus450-F and Linker-Plus275-R (sequence above) yields the full-length fragment (referred to as "Large"). A "Medium"-size fragment is amplified using Linker-Minus450-F and Linker-Plus160-R (5'-TCTTTCCCCTGCACTGTACC), and a "Small" fragment is amplified using primers Linker-Minus150-F (5'-CCTTCACCGAGG GCCTATTT) and Linker-Plus160-R. Amplification program: 1 cycle at 98 °C for 30 s, and 16 cycles of 98 °C for 10s, 63 °C for 10 s, 72 °C for 15s, and final elongation at 72 °C for 10 min.

The "Large", "Medium" and "Small" amplification products were analysed on a 0.8 % low-melt agarose gel and bands of the correct size (869 bp, 764 bp and 464 bp respectively) excised from the gel and purified using the Gel Extraction kit (QIAGEN). The vector pgRNA-pLKO.1 was cut with *Age*I (NEB), gel-purified and dephosphorylated using antarctic phosphatase (NEB). The "Large", "Medium" and "Small" amplicons were each cloned into the vector by Gibson assembly. Gibson assembly master mix was prepared as described [119], combining 320 μl 5X isothermal reaction buffer (25% PEG-8000, 500 mM Tris-HCl pH 7.5, 50 mM MgCl2, 50 mM DTT, 1 mM each of the four dNTPs and 5 mM NAD), 3 μl T5 exonuclease (10 U/μl, NEB), 20 μl of Phusion DNA polymerase (2 U/μl, NEB), 160 μl Taq DNA ligase (40 U/μl, NEB) and water in a final volume of 1.2 ml. 100ng cut vector and insert in 2-fold molar excess (total volume 5 μl) were added to 15 μl of Gibson master mix and incubated at 50 °C for 1 h. A total of 16 separate reactions were set up for each type of insert and combined for purification with the Reaction Cleanup kit (QIAGEN), followed by electroporation of the entire reaction into bacteria as described above.

2.2.4 Library QC by sequencing

A fragment 113 bp in length and comprising the gRNA targeting sequence was amplified from the library (100 ng input) using Long Amp Taq 2X Master Mix (NEB) and 1 μl of each primer gRNA-Upstream-50-F (10 μM, 5'-AAGTATT TCGATTTCTTGGCTTTATATATCT) and gRNA-Downstream-19-R (10 μM, 5'-CGGACTAGCCTTATTTTAACTTGC) in a total volume of 25 μl. Cycling conditions were 1 cycle at 94 °C for 30 s, 10 cycles of 94 °C for 30 s, 52 °C for 30 s, 65 °C for 15 s and final elongation at 65 °C for 10 min. The reaction was purified using Agencourt AMPure XP (Beckman Coulter, A63881) with a sample:bead ration of 1:2 and 5' ends were phosphorylated using T4 polynucleotide kinase (PNK, NEB). Reactions were set up as follows: 5 μl T4 DNA ligase buffer including ATP, 2 μl T4 PNK (10 U/μl) in a total reaction volume of 50 μl. Again, reactions were bead-purified using a sample:bead ratio of 1:2. Pre-annealed MisSeq adapters (Adapter-InPE-1.0: 5'- [Phos]-GATCGGAAGAGCACACGTCT, Adapter-InPE-2.0: 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATC*T) were ligated to the ends of the PCR amplicon using the Quick ligation kit (NEB) with an adapter:PCR insert ratio of 10:1 using 5 μl ligase enzyme and incubating at 18 °C for 2 h. This was followed by bead cleanup and nick translation using Long Amp Taq 2X Master Mix (NEB) in a total volume of 50 μl incubated at 72 °C for 20 min, followed by another bead cleanup. PCR amplification with KAPA HiFi PCR kit (Kapa Biosystems) with indexed reverse primers was used for indexing the sequencing libraries. Amplification conditions were as follows: 10 μl Kappa GC buffer, 1.5 μl 10 mM each dNTP mix (NEB), 2 μl 25 μM primer InPE-1.0-F (5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCT ACACGACGCTCTTCCGATC*T) and 2 μl 25 μM of indexed reverse primer (list of primers is attached below), 2.5 μl DMSO, 1 μl KAPA HiFi polymerase in a total of 50 μl. Cycling conditions were 1 cycle at 95,° C for 2 min, 6 cycles of 98 °C for 20 s, 60 °C for 15 s, 72 °C for 15 s and 1 cycle at 72 °C for 5 min.

Reactions were cleaned up using Agencourt AMPure XP (Beckman Coulter, A63881) according to the manufacturer's instructions and quantified using the Qubit dsDNA BR Assay (Life Technologies). A 4 nM library including a 5 % spike-in of PhiX control DNA was prepared for sequencing on the Illumina MiSeq according to manufacturer's instructions.

Name	Sequence
InPE-2.1-R	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGA GTTCAGACGTGTGCTCTTCCGATC*T
InPE-2.2-R	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGA GTTCAGACGTGTGCTCTTCCGATC*T
InPE-2.3-R	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGA GTTCAGACGTGTGCTCTTCCGATC*T
InPE-2.4-R	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGA GTTCAGACGTGTGCTCTTCCGATC*T
InPE-2.5-R	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGA GTTCAGACGTGTGCTCTTCCGATC*T
InPE-2.6-R	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGA GTTCAGACGTGTGCTCTTCCGATC*T
InPE-2.7-R	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGA GTTCAGACGTGTGCTCTTCCGATC*T
InPE-2.8-R	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGA GTTCAGACGTGTGCTCTTCCGATC*T
InPE-2.9-R	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGA GTTCAGACGTGTGCTCTTCCGATC*T
InPE-2.10-R	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGA GTTCAGACGTGTGCTCTTCCGATC*T
InPE-2.11-R	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGA GTTCAGACGTGTGCTCTTCCGATC*T
InPE-2.12-R	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGA GTTCAGACGTGTGCTCTTCCGATC*T

Table 2.6: PCR primers for sequencing library generation

2.2.5 Analysis of sequencing data from degenerate and MNase libraries

A detailed documentation including the code used for this analysis can be found on github at `https://github.com/annakoe/AnalysisScripts`. Briefly, gRNA sequences were extracted from the sequencing reads using cutadapt [120] and sequencing quality checked using FASTQC [121]. The number of reads per gRNA as well as the gRNA length was counted and histograms plotted using custom R code. Alignment, counting of PAM sequences, analysis of GC content and comparison to the CRISPR-Eating method [122] was performed by my collaborator Karolina Worf - the documentation for this analysis can be found at hmgubox (`https://hmgubox.helmholtz-muenchen.de:8001/d/6c6e75236e/`; password: Coralina).

2.2.6 Validation of gRNAs 30 bp and longer from the MNase library

gRNAs were selected from the sequencing output of the Lab-L sample of the MNase digest library that were 35, 40 and 45 (44, or 46) bp in length, align uniquely to the genome and are followed by a PAM sequence. These gRNA sequences were cloned into px458 (Addgene plasmid 48138), from which a gRNA can be co-expressed with a wild-type Cas9-T2A-GFP fusion protein. The vector backbone was digested with *Bbs*I (NEB), the primers pre-annealed and inserted into the backbone by Gibson cloning. Successful insertion was validated by restriction enzyme digest and the sequence validated by Sanger sequencing.

Name	Sequence
gRNA-P1-44bp-chr6-3730356-1F	TTGTGGAAAGGACGAAACACCGGAGTCTAAGC GAAGTCCCTCTCTGGGCCGGGCCTGAGACAGC AGTTTTAGAGCTAGAAATAGCAA
gRNA-P1-44bp-chr6-3730356-1R	TTGCTATTTCTAGCTCTAAAACTGCTGTCTCA GGCCCGGCCCAGAGAGGGACTTCGCTTAGACT CCGGTGTTTCGTCCTTTCCACAA
gRNA-P1-20bp-chr6-3730356-2F	TTGTGGAAAGGACGAAACACCGGGCCGGGCCT GAGACAGCAGTTTTAGAGCTAGAAATAGCAA
gRNA-P1-20bp-chr6-3730356-2R	TTGCTATTTCTAGCTCTAAAACTGCTGTCTCA GGCCCGGCCCGGTGTTTCGTCCTTTCCACAA
gRNA-H1-46bp-chr17-14203026-1F	TTGTGGAAAGGACGAAACACCGTGTTCTGTCC CCGGAACCGCTTGCAGCCAGGAGTTGGAGGCC CTCGTTTTAGAGCTAGAAATAGCAA
gRNA-H1-46bp-chr17-14203026-1R	TTGCTATTTCTAGCTCTAAAACGAGGGCCTCC AACTCCTGGCTGCAAGCGGTTCCGGGGACAGA ACACGGTGTTTCGTCCTTTCCACAA
gRNA-H1-20bp-chr17-14203026-2F	TTGTGGAAAGGACGAAACACCGCCAGGAGTTG GAGGCCCTCGTTTTAGAGCTAGAAATAGCAA
gRNA-H1-20bp-chr17-14203026-2R	TTGCTATTTCTAGCTCTAAAACGAGGGCCTCC AACTCCTGGCGGTGTTTCGTCCTTTCCACAA
gRNA-P2-40bp-chr13-53435625-1F	TTGTGGAAAGGACGAAACACCGTGAACCCAGG AGGCGGAGGTTGCAGTGAGCTGAGATCACGTT TTAGAGCTAGAAATAGCAA
gRNA-P2-40bp-chr13-53435625-1R	TTGCTATTTCTAGCTCTAAAACGTGATCTCAG CTCACTGCAACCTCCGCCTCCTGGGTTCACGG TGTTTCGTCCTTTCCACAA

gRNA-P2-20bp-chr13-53435625-2F	TTGTGGAAAGGACGAAACACCGTGCAGTGAGC TGAGATCACGTTTTAGAGCTAGAAATAGCAA
gRNA-P2-20bp-chr13-53435625-2R	TTGCTATTTCTAGCTCTAAAACGTGATCTCAG CTCACTGCACGGTGTTTCGTCCTTTCCACAA
gRNA-P3-35bp-chr10-98499080-1F	TTGTGGAAAGGACGAAACACCGTCCACCTGCC TCAGCCTCCCAAAGTGCTGGGATCGTTTTAGA GCTAGAAATAGCAA
gRNA-P3-35bp-chr10-98499080-1R	TTGCTATTTCTAGCTCTAAAACGATCCCAGCA CTTTGGGAGGCTGAGGCAGGTGGACGGTGTTT CGTCCTTTCCACAA
gRNA-P3-20bp-chr10-98499080-2F	TTGTGGAAAGGACGAAACACCGCTCCCAAAGT GCTGGGATCGTTTTAGAGCTAGAAATAGCAA
gRNA-P3-20bp-chr10-98499080-2R	TTGCTATTTCTAGCTCTAAAACGATCCCAGCA CTTTGGGAGCGGTGTTTCGTCCTTTCCACAA
gRNA-Z1-35bp-chr19-37340687-1F	TTGTGGAAAGGACGAAACACCGGAGTGTGTGG AGGTGGGGGGGGCAAGATCGTGTAGTTTTAG AGCTAGAAATAGCAA
gRNA-Z1-35bp-chr19-37340687-1R	TTGCTATTTCTAGCTCTAAAACTACACGATCT TGCCCCCCCCACCTCCACACACTCCGGTGTTT CGTCCTTTCCACAA
gRNA-Z1-20bp-chr19-37340687-2F	TTGTGGAAAGGACGAAACACCGGGGGGGGGCA AGATCGTGTAGTTTTAGAGCTAGAAATAGCAA
gRNA-Z1-20bp-chr19-37340687-2R	TTGCTATTTCTAGCTCTAAAACTACACGATCT TGCCCCCCCCCGGTGTTTCGTCCTTTCCACAA

Table 2.7: Oligonucleotides for Gibson cloning of selected gRNAs greater than 30 bp in length from the human MNase digest library

The vector px458 harboring the different gRNAs was transiently transfected into HEK293T cells (ATCC 293T/17, CRL-11268) grown in DMEM (Life Technologies) supplemented with 10 % FBS (Life Technologies,). Cells were transfected using Lipofectamine (Life Technologies 15338100). 5x10^5 cells were seeded per well in 6 well plates the day before transfection. 2.5 μg vector DNA was added to 500 μl Optimem medium and mixed, followed by addition of 5 μl Lipofectamine LTX reagent, mixing and incubation at room temperature for 30 min before addition to cells in 2 ml fresh medium. Media was changed the next day and cells harvested 48 hours after transfection. DNA was extracted using the DNeasy blood and tissue kit (QIAGEN) according to the manufacturer's instructions.

gRNA target regions were amplified from the samples and controls by PCR using the primers in **Table 2.8**.

The PCR reactions were set up using 2X Phusion High-Fidelity PCR Master Mix with GC Buffer (NEB) in a 50 μl reaction with 2 μl of each primer at a dilution of 10 μM and adding 100 ng input DNA. The annealing temperatures for primers amplifying each of the target regions was optimized and found to be 68 °C for target of gRNA P1, 65 °C for target of gRNA H1, 64 °C for target of gRNA P2, 66 °C for target of gRNA P3, and 68 °C for target of gRNA Z1.

Amplification program: 1 cycle at 98 °C for 30 s, 12 cycles of 98 °C for 10s, optimised annealing temperature as above for 10 s, 72 °C for 10s, and final elongation at 72 °C for 10 min. Reactions were cleaned up using Agencourt Ampure beads (Beckman Coulter) at a sample:beads ratio of 1:1.5. In order to prepare libraries for sequencing on the Illumina MiSeq platform, Illumina Nextera indexed adapter sequences were added in a second round of PCR amplification using KAPA HiFi PCR kit (Kapa Biosystems) in a 25 μl reaction with Kapa GC Buffer, 0.75 μl dNTP mix, 0.5 μl KAPA HiFi polymerase, 0.75 μl i5-indexed forward primer (10 μM) and 0.75 μl i7-indexed reverse primer (10 μM) and half the cleaned-up product of the first round of amplification. Cycling conditions were 1 cycle at 95,° C for 5 min, 6 cycles of 98 °C for 20 s, 60 °C for 15 s, 72 °C for 30 s and 1 cycle at 72 °C for 5 min. Products were purified using Agencourt Ampure beads (Beckman Coulter) at a sample:beads ratio of 1:0.8. Libraries were quantified using the Quant-iT PicoGreen dsDNA assay (Thermo Fisher, P11496) and by qPCR using the Kapa Library Quantification kit (Roche, 07960140001). Libraries were sequenced on the Illumina MiSeq platform. Data analysis was performed using a custom pipeline written by Javier Herrero (available through github: `https://github.com/jherrero/crispr-parsr/tree/bd255a125192b474a738ec01e7ce2c0428bcd10a`). The pipeline was run using standard parameters except for addition of the "allow-any" flag.

Name	Sequence
Target-gRNA-P1-1F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAATCCTTC CTTAATTGCCTGTGAC
Target-gRNA-P1-1R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAGCTT CTCAGGGACCATCTTTAG
Target-gRNA-H1-1F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCGGTGA GTCACTTCGTGAG
Target-gRNA-H1-1R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCAAAT TCTTACTGGTCGTGTTCA
Target-gRNA-P2-1F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCCCTCAT CTCTTCTAACCATCAG
Target-gRNA-P2-1R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAACAT GGTGAAATCCCATATCTAC
Target-gRNA-P3-1F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCCACGCC TAGCTACATTTTTG
Target-gRNA-P3-1R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAAGGA AGGAAGGAAGGAGGAA
Target-gRNA-Z1-1F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGAGAC TTGGAGGTTCAGATTCC
Target-gRNA-Z1-1R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTCAGA TTGGTGTGTGTGTGAGAG

Table 2.8: PCR primers for amplification of target regions of selected gRNAs greater than 30 bp in length from the human MNase digest library

2.2.7 Construction of the EMT5000 library

A small library containing gRNAs tiling along the promoters of 15 known EMT genes was designed as follows (detailed documentation and code available on github at `https://github.com/annakoe/AnalysisScripts`): Briefly, target regions were defined by inspection of the epigenetic marks around the promoters of a list of genes from ref. [123]. The following genomic regions where chosen as target sites for the design of gRNAs:

Chromosome	Start	Stop	Target gene
chr1	170626538	170637878	PRRX1
chr2	145272896	145282545	ZEB2
chr2	145310788	145311630	ZEB2
chr6	166578775	166584033	Brachyury (T gene)
chr6	166586466	166588249	Brachyury (T gene)
chr7	19155427	19162115	TWIST1
chr8	49831094	49838789	SLUG
chr10	31549929	31552360	ZEB1
chr10	31603262	31611019	ZEB1
chr14	61113258	61126351	SIX1
chr14	95235188	95236645	GSC (Goosecoid)
chr16	68765472	68768900	Cdh1
chr16	68770501	68779468	Cdh1
chr16	86596822	86601033	FOXC2
chr18	25616470	25616815	Cdh2
chr18	25753143	25759002	Cdh2
chr18	25763319	25764152	Cdh2
chr18	25783828	25784775	Cdh2
chr18	52966479	52970132	TCF4
chr18	52983584	52991765	TCF4
chr18	52994740	52997201	TCF4
chr18	53067559	53071402	TCF4
chr18	53072594	53073776	TCF4
chr18	53087724	53090359	TCF4
chr18	53176467	53178784	TCF4
chr18	53252816	53257791	TCF4
chr18	53259747	53260381	TCF4
chr18	53301547	53303603	TCF4
chr19	1631468	1633670	E47/TCF3
chr19	1646514	1653855	E47/TCF3
chr19	1655335	1656204	E47/TCF3
chr19	1660918	1661677	E47/TCF3
chr20	48592707	48600991	SNAIL1
chrX	56258091	56260688	KLF8

Table 2.9: EMT5000 library target regions

All gRNAs that are followed by an NGG PAM sequence and fall into the above re-

gions were then retrieved by searching for the sequences $GN_{20}GG$ and only those that align uniquely were selected for synthesis. A pool of oligonucleotides of the form 5'-TCTTGTGGAAAGGACGAAACACC-GN19-GTTTTAGAGCTAGAA ATAGCAAGTTAAAATAAGGCT-3', where GN_{19} donates the 5,086 different guide sequences, was obtained from Custom Array Inc. The sequence was then amplified using 0.5 μM of each primer TSS-upstream (CTTGTGGAAAGGAC-GAAACA) and TSS-downstream (GCCTTATTTTAACTTGCTATTTCTAGC) and 1 ng of the custom oligonucleotide pool as input in a 25 μl reaction with 2X Phusion HiFi GC Master Mix. Cycling conditions were as follows: 1 cycle at 98 °C for 30 s, 30 cycles of 98 °C for 10 s, 59 °C for 10 s, 72 °C for 10 s and final elongation at 72 °C for 10 min. A total of 3 PCR reactions were set up for this library prep. Reactions were purified using the MinElute reaction cleanup kit (QIAGEN).

The vector pgRNA-pLKO.1 was digested with *Age*I (NEB) and linearised vector was gel-purified from a 0.8 % low-melt agarose gel using the QIAQUICK gel extraction kit (QIAGEN). 2 μg of purified linearised vector was dephosphorylated by treatment with 3U of shrimp alkaline phosphatase (NEB) at 37 °C for 40 min, followed by reaction cleanup using the QIAQUICK reaction cleanup kit (QIAGEN). Gibson reactions were set up using 0.08 pmol vector and 0.42 pmol PCR amplicon in a 20 μl reaction with Gibson master mix and incubated at 50 °C for 60 min. A total of 5 Gibson reactions were set up.

The Gibson master mix was prepared by adding 0.64 μl of 10 U/μl T5 exonuclease (NEB), 20 μl of 2 U/μl Phusion DNA polymerase (NEB), 160 μl of 40 U/μl Taq DNA ligase to 320 μl 5X isothermal reaction buffer (3 ml of 1 M Tris-HCl pH 7.5, 300 μl of 1 M MgCl2, 60 μl of 100 mM GTP, 60 μl of 100 mM dATP, 60 μl of 100 mM dTTP, 60 μl of 100 mM dCTP, 300 μl of 1 M DTT, 1.5 g PEG-8000 and 300 μl of 100 mM NAD in a total volume of 6 ml), and bringing the total volume up to 1.2 ml with nuclease-free water.

Gibson reactions were cleaned up using the Qiaquick PCR purification kit (QI-AGEN), eluting twice in 15 μl water. The entire library was electroporated into home-made electro-competent *E.coli* cells as described above and allowed to re-cover in non-selective 2TY medium at 37 °C for 1 h. For library QC $1/1000^{th}$ of the total ligation was plated out on a 10 cm petri dish, DNA extracted from 20 single colonies using the QIAQUICK Spin Miniprep kit (Qiagen) and gRNAs sequenced by Sanger sequencing. The remainder of the electroporated library was plated onto 2TY-coated plates (Bio-assay dish with lid, 245mm x 245mm x 25mm,

radiation sterilized, Thermo Scientific Nunc) containing 100 μg/μl Ampicillin. Following overnight incubation at 37 °C, the bacteria were harvested by scraping and the plasmid library extracted using the HiSpeed Plasmid Maxi Kit (QIAGEN). QC for the EMT5000 control library was conducted by Sanger sequencing. Individual colonies were picked from a separate 10cm petri dish, plasmid DNA extracted using the Spin Miniprep Kit (Qiagen), and 19 clones sequenced.

2.2.8 Packaging of libraries based on gRNA-pLKO.1 into lentiviral particles

Lentiviral packaging was carried out according to standard procedures by Catherine King, who provides this service to the Cancer Genome Engineering Facility (CAGE) as described in section 2.1.7

2.2.9 Determination of viable virus titer in A549 cells

0.5x10^5 A549 cells were plated in 12-well plates. The next day, cells were transduced with virus at various concentrations in complete media containing with or without 10 μg/ml Polybrene (Millipore). The next day, cells were washed and trypsinised and each well split into two wells containing equal amounts of cells. Cells in one of the replica wells were plated in medium containing 1.5 μg/ml Puromycin (Gibco). On day 3 after transduction, cells in each of the wells were trypsinised and counted. Only wells containing <20 % of viable cells were used for the calculation of viable titer according to the following formula:

Viable titer in transducing units/ml (TU/ml) = % antibiotic positive cells x number of cells transduced/vol of virus added to well (ml)

were % cells transduced = number of cells in well + puromycin/ number of cells in corresponding well - Puromycin

2.3 Preliminary screen with EMT5000 library in A549 cells

7.5x10^5 A549 cells were seeded per 10 cm dish on day 1 and transduced with the lentiviral EMT5000 library at an multiplicity of infection (MOI) of 0.3 the

next day in the presence of Polybrene (Millipore) at a concentration of 10 μg/ml. Medium was changed the next day and selection started with the addition of 1.5 μg/ml Puromycin (Gibco). On day 5, cells were trypsinised. For the positive control, 1 x 10^5 cells transduced with the EMT5000 library were plated in 6-well plates with or without StemXVivo EMT Inducing Media Supplement (R& D systems). Medium was changed every 3 days thereafter.

The successfully transduced cells were then transfected using the fast-forward transfection method with Lipofectamine LTX (Invitrogen) on day 5. Transfection mixes were made up by pre-mixing 4-10 μg Cas9-chromatin modifier vector (1.4 pmol) with 10 μl Plus reagent in 1 ml OptiMEM (Gibco), and 40 μl Lipofectamine LTX in 1 ml OptiMEM. The DNA mix was added to the Lipofectamine dilution and incubated for 20 min at room temperature. Cells were plated at a density of 3x10^6 cells per 10 cm dish and transfection mixes immediately added to the cell suspension. Medium was changed the next day and selection with 2 mg/ml Hygromycin (Life Technologies) started 48 hours after transfection. On day 10, cells were detached using Versene (Gibco) and prepared for FACS sorting.

2.3.1 Preliminary Screen - Staining and FACS sorting

Following detachment, cells were washed in sorting buffer (1 % FBS in PBS) and stained with Fixable Viability Dye 450 (Beckman Dickson) at a dilution of 1:1000 in PBS at 4 °C for 30 min protected from light. Cells were washed twice with sorting buffer, and stained with 80 μg anti E-Cadherin-FITC (CD324, clone 67A4, cat. A15757, Life Technologies) and 100 μg N-Cadherin-PE (CD325, clone 8C11, cat. 561554, BD Biosciences) per 100 μl cell suspension containing 5x10^5 cells for 30 min at 4 °C away from light. Cells were washed twice with sorting buffer and fixed with 4 % formaldehyde solution (Merck) for 10 min at room temperature. Cells were washed twice with sorting buffer and pipetted through a 70 μm cell strainer (BD Falcon). Cells were sorted on a FACS ARIA III (BD Biosciences).

2.4 Screen with the EMT5000 control library

Polyclonal or monoclonal stable A549 cell lines were maintained in Ham's F-12K (Kaighn's) Medium (Gibco) with 10 % FBS with 30 μg/ml Blasticidin (Life

Technologies). 4x10^5 cells were plated per T75 flask the day before transduction. The next day, cells were transduced with EMT5000 lentivirus in medium without Blasticidin at an MOI of <0.3 to ensure single insertions per cell. Addition of polybrene, a commonly used transduction enhancer, was omitted here as suggested previously [124]. Polybrene is a cationic polymer that facilitates virus entry into cells by neutralizing the charge between the virus envelope and the cell membrane. However, it has also been shown to increase virus aggregation which in turn increases the possibility of multiple virus infection per cell in a pooled screen (although this effect may be relatively small [125]).

12 hours after transduction, selection was started with 1.5 μg/ml Puromycin (Life Technologies). Cells were maintained in Puromycin selection and split using Versene (Gibco) for detachment to preserve surface markers. For the positive control, 2x10^5 untransduced control cells were plated onto a T75 with StemXVivo EMT Inducing Media Supplement (R& D systems) on day 3 of the experiment (1 day after virus transduction). On day 7 after transduction, cells were harvested for sorting. Cells were detached with Versene (Gibco) and washed once with cold PBS (Gibco). Cells were counted and subsequently stained with Fixable Viability Stain 450 (BD Biosciences) diluted in PBS at a concentration of 5x10^6 cells/ml for 30 min at 4 °C away from light. Cells were washed twice (1x PBS, 1x Sorting Buffer - 1% FBS in PBS) and stained with 160 μg anti E-Cadherin-FITC (CD324, clone 67A4, cat. A15757, Life Technologies) and 100 μg N-Cadherin-PE (CD325, clone 8C11, cat. 561554, BD Biosciences) diluted in sorting buffer per 100 μl cell suspension containing 5x10^5 cells for 30 min at 4 °C away from light. Cells were washed twice with sorting buffer and fixed in 4 % formaldehyde solution (Merck) for 5 min at room temperature. Cells were washed twice with sorting buffer and pipetted through a 70 μm cell strainer (BD Falcon). Cells were sorted on a FACS ARIA III (BD Biosciences) into sorting buffer.

2.4.1 Extraction of DNA from sorted cells

Cells were sorted directly into Quick Extract DNA extraction solution (Epicentre). Tubes were vortexed for 15s, incubated at 65 °C for 6 min, vortexed for 15s and incubated at 98 °C for 2 min before storage at -20 °C.

2.4.2 Amplification and sequencing of gRNA sequences from sorted cells

Option A: 12.5 μl Phusion High-Fidelity HS Master Mix (NEB), 1 of each μl 10 μM barcoded primer (NexteraXT-N7-50-1F, NexteraXT-N7+19-1R) and 10.5 μl of template (cell lysate in Quick Extract solution). These reactions were incubated: 1 cycle 98 °C for 30s, 2 cycles 98 °C for 10s, 60 °C for 10s, 72 °C for 10s, 72 °C for 10 min followed addition of 9.8 μl ExoSAP-IT and incubation at 37 °C for 30 min and inactivation of ExoSAP-IT at 80 °C for 15 min. This was followed by a second round of amplification by adding 15 μl Phusion High-Fidelity HS Master Mix (NEB) and 1 μl each of 10 μM primers Nextera 2F/2R (Nextera 2F/2R yielded a nonspecific amplicon, which was removed using agarose gel excision and purification) or Nextera 3F/3R respectively. 1 cycle 98 °C for 30s, 28 cycles 98 °C for 10s, 60 °C for 10s, 72 °C for 10s, 72 °C for 10 min.

Option B: Several 25 μl PCR reactions were set up per sample as follows: 12.5 μl Phusion High-Fidelity HS Master Mix (NEB), 1 of each μl 10 μM barcoded primer (NexteraXT-N7-50-1F, NexteraXT-N7+19-1R) and 10.5 μl of template (cell lysate in Quick Extract solution). These reactions were incubated: 1 cycle 98 °C for 30s, 3 cycles 98 °C for 10s, 60 °C for 10s, 72 °C for 10s, 72 °C for 10 min followed by addition of 150 ng pEGFP-C1 plasmid as carrier DNA and column purification using the PCR reaction cleanup kit (QIAGEN), followed by bead purification with Ampure XP (Beckman Coulter) at a sample:bead ration of 1:0.8 to remove primer dimers. This was followed by a second round of PCR with 12.5 μl Phusion High-Fidelity HS Master Mix (NEB), 1 of each μl 10 μM primer (NexteraXT-3F, NexteraXT-3R). Cycling conditions were: 1 cycle 98 °C for 30s, 26-32 cycles 98 °C for 10s, 65 °C for 10s, 72 °C for 10s, 72 °C for 10 min. This was followed by a reverse bead cleanup and agarose gel excision cleanup to remove carrier DNA.

Next, barcoded Nextera sequencing adapters (Illumina) were added to the fragments. 5 μl 5X Kapa GC Buffer, 0.75 μl dNTPs, 0.5 μl Kapa HiFi polymerase, 0.75 μl Nextera i5F (10 μM) , 0.75 μl Nextera i7R (10 μM) and 2 μl template ($1/5^{th}$ of purified PCR reaction from above) in a total volume of 25 μl. Cycling conditions were: 1 cycle 95 °C for 5 min, 6 cycles 98 °C for 20s, 65 °C for 15s, 72 °C for 15s, 72 °C for 10 min.

Libraries were sequenced on a HiSeq2000 using TruSeq Dual Index Sequencing

Name	Sequence
NexteraXT-N7-50-1F (1^{st} round)	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GNNNNNNNAAGTATTTCGATTTCTTGGCTT
NexteraXT-N7+19-1R (1^{st} round)	GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC AGNNNNNNNCGGACTAGCCTTATTTTAACTTG
NexteraXT-2F (2^{nd} round, non-specific)	TCGTCGGCAGCGTC
NexteraXT-2R (2^{nd} round, non-specific)	GTCTCGTGGGCTCGG
NexteraXT-3F (2^{nd} round, specific)	TCGTCGGCAGCGTCAGATGT
NexteraXT-3R (2^{nd} round, specific)	GTCTCGTGGGCTCGGAGATGTG

Table 2.10: PCR primers for making the degenerate gRNA library using circle amplification

Primers, with the TruSeq SBS Kit v3 - HS (200 Cycles) reagents in Fast mode (single read).

2.5 Data analysis - Screen with the EMT5000 control library

A detailed documentation including the code used for this analysis can be found on github at `https://github.com/annakoe/AnalysisScripts`. First, the sequence of the gRNA and the 5' and 3' barcode, which serve as unique molecular identifiers (UMIs), were extracted from each read using cutadapt (version 1.2.1) [120]. Reads with any base having a sequencing quality score of < 20 in these regions were discarded. The gRNA sequence was aligned back to the reference EMT5000 library with bwa (version: 0.6.2-r126) [126], allowing a maximum of 2 mismatches and 1 gap opening. Only sequences that mapped uniquely to the forward strand were extracted. The read ID, gRNA coordinates (chr:start-stop) and 5' and 3' barcodes (joined to give a 14 bp unique molecular identifier) were used as input for a Bayesian model for PCR error correction. This model (contributed by James E. Barrett) infers the number of unique barcoded gRNA molecules from noise-corrupted count data. Following PCR error correction the gRNA sequences and associated UMI-corrected counts were fed into DESeq2 [127] to identify gRNA sequences enriched in the samples relative to the negative controls. Next, Log2Fold Changes (and associated standard errors) calculated for each gRNA using DESeq2 were used to rank gRNAs and identify

those that display consistent positive fold change across the different screening experiments. Ranking was performed using an R package called "DesiR" [128], which implements desirability functions for ranking and prioritising candidates in a variety of settings. gRNAs were ranked based on positive Log2FoldChange and small Log2FoldChange standard error across the different screening experiments, whereby Log2FoldChange was given four times the weight of Log2FoldChange standard error and all experiments were weighted equally. The top 10 ranked gRNAs were chosen as candidates for validation.

2.6 Validation of candidate gRNAs

Candidate gRNAs were cloned into the lenti-gRNA-pLKO.1 vector using circle amplification. PCR reactions were set up as follows: 12.5 μl Phusion HF Master mix with GC buffer (NEB, M0532S), 1 μl of 10 μM reverse primer MLM3636-gRNA-R, 1 μl of 10 μM forward primer (see list below) and 1 ng of plasmid template either lenti-gRNA-pLKO.1 my-gRNA-Neo.

Primer sequences were:

Name	Sequence
MLM3636-gRNA-R	5' [P]-CGGTGTTTCGTCCTTTCCAC
Set7-C1-chr6:166583296-166583319-F	5' [P]-GCCTCGGAACCCTAGGCACGTTTTAG AGCTAGAAATAGCAAGTTAAAATAAGG
Set7-C2-p300-C2-chr10:31609152-31609175-F	5' [P]-CCGGGAGCCGCGCGGATGGGTTTTAG AGCTAGAAATAGCAAGTTAAAATAAGG
Set7-C3-chr19:1652696-1652719-F	5' [P]-CGGATCCCTCCGCCACCTCGTTTTAGA GCTAGAAATAGCAAGTTAAAATAAGG
Set7-C4-chr20:48599059-48599082-F	5' [P]-GAAATTTCCTCCGCCCGGCGTTTTAGA GCTAGAAATAGCAAGTTAAAATAAGG
Set7-C5-chr10:31609192-31609215-F	5' [P]-TGTTTGCGGAGTTGTTACCGTTTTAG AGCTAGAAATAGCAAGTTAAAATAAGG
Set7-C6-chr18:53303443-53303466-F	5' [P]-GCAGAGCAGGCTGGTTTTTGTTTTAG AGCTAGAAATAGCAAGTTAAAATAAGG
Set7-C7-chr6:166583988-166584011-F	5' [P]-TTTCAAGCATGTTTTCGGTGTTTTAGA GCTAGAAATAGCAAGTTAAAATAAGG
Set7-C8-chr19:1631739-1631762-F	5' [P]-CAGGGCTCCTCCTGCTTCCGTTTTAGA GCTAGAAATAGCAAGTTAAAATAAGG

Set7-C9-chr19:1649153-1649176-F	5' [P]-ACTGCCACGGTTATCACTGGTTTTAGA GCTAGAAATAGCAAGTTAAAATAAGG
Set7-C10-chr14:61117798-61117821-F	5' [P]-TGCTGGGCTCTTTGGCTAAGTTTTAG AGCTAGAAATAGCAAGTTAAAATAAGG
p300-C1-chr14:95236066-95236089-F	5' [P]-CACGTGCAGGCGGCGCCCGGTTTTAG AGCTAGAAATAGCAAGTTAAAATAAGG
p300-C3-chr19:1655867-1655890-F	5' [P]-CCAAGGACAACTTCCCACTGTTTTAGA GCTAGAAATAGCAAGTTAAAATAAGG
p300-C4-chr18:53255433-53255456-F	5' [P]-TGTTAAGAGTCAGGGATCTGTTTTAG AGCTAGAAATAGCAAGTTAAAATAAGG
p300-C5-chr6:166581382-166581405-F	5' [P]-GCAGCGCTGGGGTGCTCGGGTTTTAG AGCTAGAAATAGCAAGTTAAAATAAGG
p300-C6-chr16:68775128-68775151-F	5' [P]-TTTTGGCTTTTTTGGACTGGTTTTAGA GCTAGAAATAGCAAGTTAAAATAAGG
p300-C7-chr18:53088256-53088279-F	5' [P]-TGAAATTCTACCATCTGGAGTTTTAGA GCTAGAAATAGCAAGTTAAAATAAGG
p300-C8-chr14:95235538-95235561-F	5' [P]-TCGCAGGCTACGAGGGCCCGTTTTAG AGCTAGAAATAGCAAGTTAAAATAAGG
p300-C9-chr18:52988832-52988855-F	5' [P]-TCGCAGGGATGAGCATCCTGTTTTAG AGCTAGAAATAGCAAGTTAAAATAAGG
p300-C10-chr14:95236264-95236287-F	5' [P]-TTGCCGGTGGCGCACAGCGGTTTTAG AGCTAGAAATAGCAAGTTAAAATAAGG
Neg-C1-GAPDH-F	5' [P]-GGTGGAGTCGCGTGTGGCGGTTTTAG AGCTAGAAATAGCAAGTTAAAATAAGG
Neg-C2-scramble-F	5' [P]-GGCGCTCCACGCGATACCAGTTTTAG AGCTAGAAATAGCAAGTTAAAATAAGG

Table 2.11: Primers for cloning candidate gRNAs

Cycling conditions were: 1 cycle 98 °C for 30s, 28 cycles 98 °C for 10s, 62 °C for 10s, 72 °C for 3.5 min, followed by 72 °C for 10 min.

This was followed by addition of 1 μl T4 DNA polymerase (NEB) and incubation at 12 °C for 20 min and bead purification with Ampure XP beads following the manufacturer's instructions (1:1 sample:bead ratio). PCR products were

eluted in 30 μl H_20 and subjected to ligation under dilute conditions to promote intramolecular ligation. Ligations were set up as follows: 30 μl purified PCR reaction, 10 μl 10X DNA ligase buffer, 1.3 μl concentrated T4 DNA ligase (NEB) and water up to 100 μl. Reactions were incubated at 16 °C for 16 hours. This was followed by digestion with *Dpn*I to remove circular plasmid template. To the ligation reactions 11.4 μl of Cut Smart Buffer and 4 μl *Dpn*I (20 U/μl, NEB) were added, followed by incubation at 37 °C for 2 hours, followed by cleanup using QIAGEN PCR reaction cleanup kit. Reactions were transformed into Top10 chemically competent bacteria (Invitrogen). The constructs were validated by restriction enzyme digest and Sanger sequencing.

2.6.1 Packaging of candidate gRNAs in lenti-gRNA-pLKO.1 backbone into lentivirus and virus concentration by PEG precipitation

Lentivirus production was performed under GM number UCL RA002231/1 (HSE GM 14/15.3). HEK293T (ATCC 293T/17, CRL-11268) cells were maintained in complete DMEM (Life Technologies, 41966052) with 10% FBS. Cells were plated at a density of $2.5x10^6$ cells per 10 cm dish and transiently transfected with the lentiviral expression construct and the two packaging vectors p8.91 (gag-pol expressor) and pMDG (VSV-G expressor, generous gift of Catherine King) using Fugene transfection reagent (Promega, E2691) the next day. Medium was changed the following day. Lentivirus-containing medium was harvested 24 and 48 hours after this initial media change. The harvested medium was filtered through a 0.22 μm filter.

Virus-containing medium was further concentrated by PEG precipitation. 5X PEG solution was prepared by dissolving 200g PEG, 12g NaCL, 1ml of 1M Tris pH7.5 in 500 ml H_20, and adjusting the pH to 7.2, followed by sterile filtration. Equal volumes of 5X PEG and virus-containing medium were mixed and incubated at 4 °C overnight before centrifugation at 1,500 g for 30 min at 4 °C. Supernatant was removed an the virus pellet resuspended in PBS and stored in aliquots at -80 °C until further use.

2.6.2 Validation experiment

A549 cells expressing a dCas9-chromatin modifier fusion protein were maintained in Ham's F-12K (Kaighn's) Medium (Gibco) with 10 % FBS and 1 % Pen/Strep (Gibco) with 30 μg/ml Blasticidin (Life Technologies). 0.7x10^5 cells were plated in 6-well plates and transduced with virus containing candidate gRNAs the next day. 24 hours after transduction selection with 1.5 μg/ml Puromycin was started. On day 5 after transduction cells were detached using Versene (Gibco), counted and washed in PBS before staining with Fixable Viability Dye 450 (Beckman Dickson) at a dilution of 1:1000 in PBS at 4 °C for 30 min protected from light. Cells were washed twice with sorting buffer, and stained with 80 μg anti E-Cadherin-FITC (CD324, clone 67A4, cat. A15757, Life Technologies) and 100 μg N-Cadherin-PE (CD325, clone 8C11, cat. 561554, BD Biosciences) per 100 μl cell suspension containing 5x10^5 cells for 30 min at 4 °C away from light. Cells were again washed twice with sorting buffer, filtered through a 70 μm cell strainer (BD Falcon) and analysed by flow cytometry on a Fortessa X20 instrument. Data were analysed using FlowJo software.

2.6.3 Extraction of DNA and RNA from monoclonal stable cell lines (dCas9-p300 and dCas9-SET7 constructs)

A549 cells were maintained in Ham's F-12K (Kaighn's) Medium (Gibco) with 10 % FBS (Gibco). Monoclonal stable cell lines were further maintained under selection with 30 μg/ml Blasticidin (Life Technologies). Genomic DNA was extracted from around 3x10^6 cells using the DNeasy Blood and Tissue kit (QIAGEN). Total RNA was extracted from around 2x10^6 cells using the RNeasy Mini kit (QIAGEN) according to the manufacturer's instructions including a DNase digestion step.

2.6.4 Amplification of the dCas9-chromatin modifier-T2A-blasticidin expression cassette from genomic DNA

PCR reactions were set up using 2X Phusion HotStart Master Mix (NEB), 0.5 mM of each primer and 50 ng genomic DNA in a 25 μl reaction. Primer sequences can be found in **Table 2.12** below. Cycling conditions were: 1 cycle 98 °C for 30s, 35 cycles 98 °C for 10s, 58 °C for 10s, 72 °C for 30 min, followed by 72 °C for

10 min, except for reactions using primers C-1F/C-1R and C-2F/C-1R, where an extension time of 1 min was used. Reactions were analysed on a 1 % agarose gel.

Name	Sequence
prom-1F	TGGAATTTGCCCTTTTTGAG
prom-1R	TGGCAGCCAAAAATAAGTCC
C-1F	GCAGCTCCTAAATGCGAAAC
C-2F	ACCAATCCATCACGGGATTA
C-1R	CAACACCACGGAATTGTCAG
C-3F	GCACAATTACCCGGAGAGAA
C-3R	TGTGACGTCCCATGACCTTA
C-4F	TGAATGCTTCGATTCTGTCG
C-4R	TGTCACTTTTCCCTCGGTTC
C-5F	GATGCCATTGTACCCCAATC
C-5R	CCTTTTTACGAGCGATGAGC
C-6F	GCGAACAGGAGATAGGCAAG
P-6R	AGAGGGCTTTGGTTCGGTAT
Set-7F	ACCAATCCATCACGGGATTA
Set-7R	AGACTTCCTCTGCCCTCTCC
Set-8F	GATGGGGAGATGACTGGAGA
Set-8R	TGGGGATGCTGTTGATTGTA
P-7F	GATCCCCCAAGAAGAAGAGG
P-7R	GCTGTCTCCCTTGGTCACAT
P-8F	GCCCAATGTTCTGGAAGAAA
P-8R	AGCAATTCACGAATCCCAAC
Blast-9F	TTTTACTGGGGGACCTTGTG
Blast-9R	GAGATCCGACTCGTCTGAGG

Table 2.12: Primers for amplification of the integrated dCas9-chromatin modifier-T2A-blasticidin expression cassette from genomic DNA

Chapter 3

Results: Establishing a CRISPR-based epigenetic screening method

3.1 dCas9-chromatin modifier constructs

3.1.1 Construction and transient expression in HEK293T cells

The known catalytic domains of several chromatin-modifying enzymes (**Figure 3.1**) were identified from the references included below. The catalytic domain of the following enzymes were each cloned into a vector downstream of dCas9, separated by a Flag-tag and a G3S-linker: the DNA demethylase TET1 [129], the DNA methyltransferase DNMT3a [79], the H3K4 histone methyltransferase SET7 [130], LSD1 [131], which demethylates both mono- and di-methylated H3K4 as well as H3K9, the H3K9 methyltransferase G9a [132], JMJD2A [133], which demethylates H3K9me3 and H3K36me3, as well as p300 [134], which acetylates all four core histones including H3K122 and H3K27, and HDAC1 [135], which removes these acetylation marks. Successful construction of these plasmids was verified by Sanger sequencing (for vector maps see Appendix).

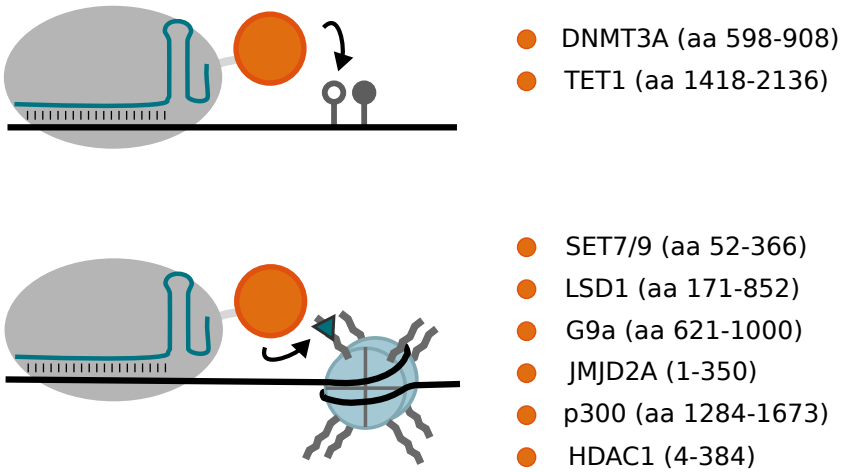


Figure 3.1: Schematic illustration of dCas9-chromatin modifier fusion constructs. These can be used to add or remove DNA methylation as well as histone marks as indicated.

Expression of the dCas9-chromatin modifier fusion proteins in HEK293T cells was ascertained using Western blotting (**Figure 3.2 A**) following transient transfection (without selection). All fusion proteins are detectable in whole cell extracts. The constructs encode large proteins - dCas9 alone is around 160 kDa - and the fusion proteins with the attached chromatin modifier are between 168 and 241 kDa in size. The additional lower molecular weight bands visible below the expected band probably represent degradation products or perhaps proteins for which transcription or translation terminated early.

Nuclear and Cytoplasmic fractionation showed that most of the constructs can be detected in the nucleus (**Figure 3.3 A**); they appear to be imported despite their large size. Staining with MEK1, which is mainly cytoplasmic shows that there is hardly any contamination of the nuclear fraction with cytoplasmic protein. Conversely, p53 is present mostly in the nucleus, as expected (**Figure 3.3 A**).

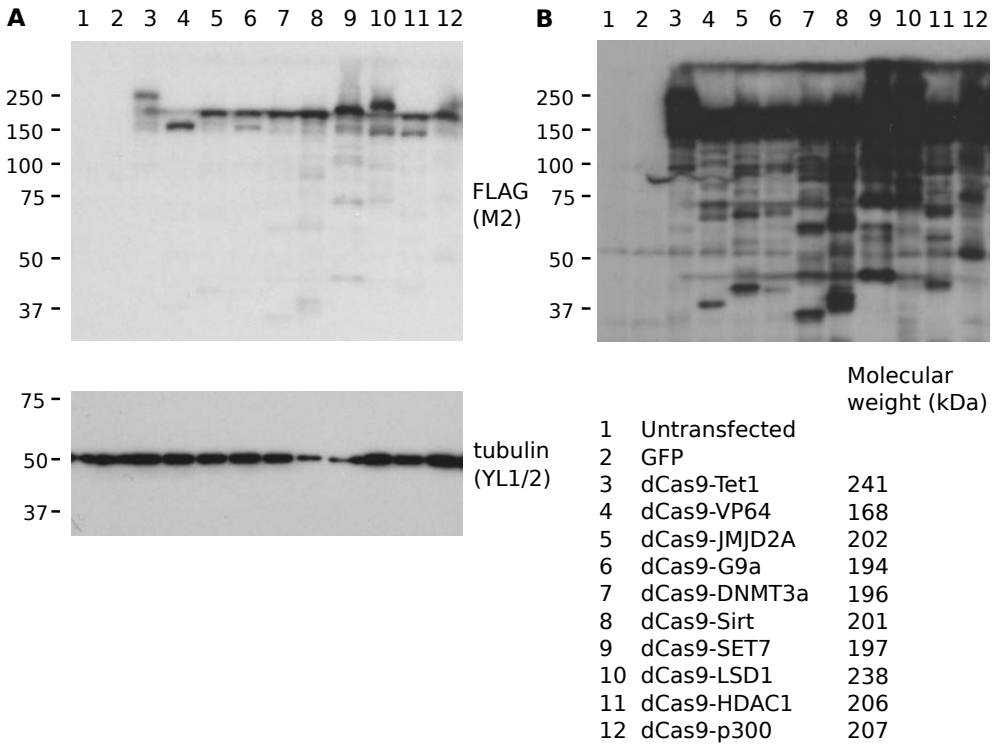


Figure 3.2: Expression of Cas9-chromatin modifier constructs in HEK293T cells. A. Western blotting using anti-Flag M2 antibody against Cas9-chromatin modifier constructs in whole-cell extracts. Tubulin was used as a loading control. **B.** Longer exposure of the blot.

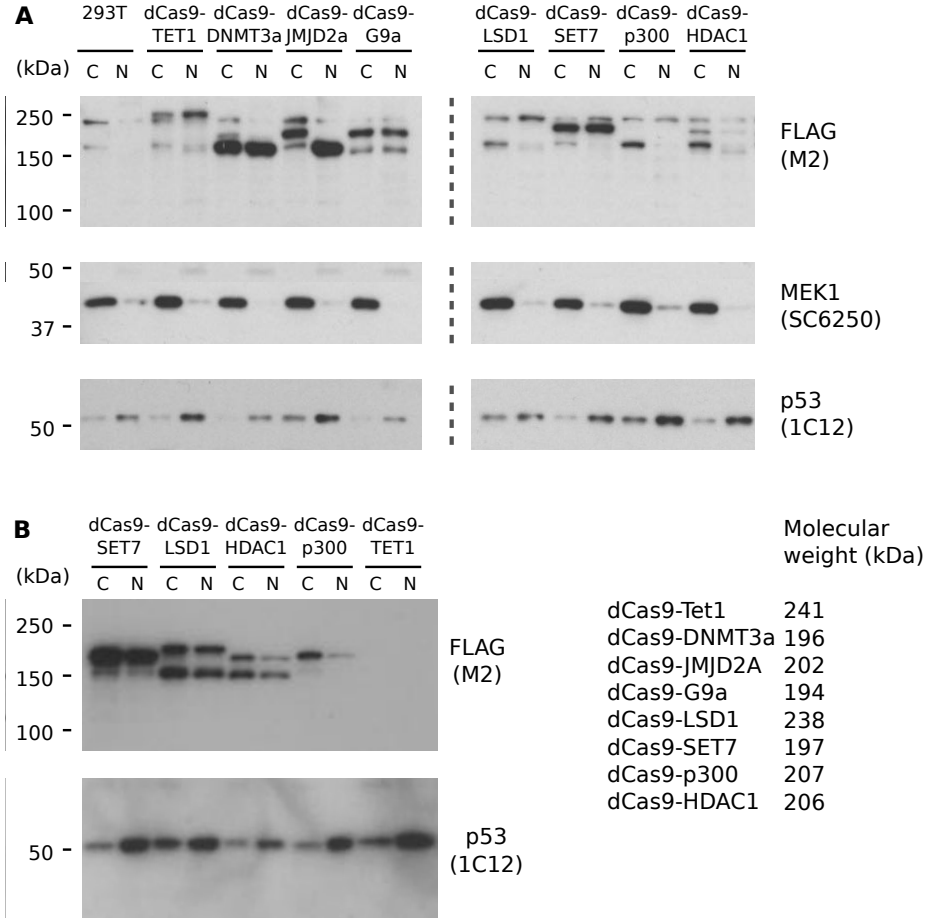


Figure 3.3: Expression of dCas9-chromatin modifier constructs in nuclear and cytoplasmic fractions from HEK293T cells. A. Western blotting using anti-Flag M2 antibody against dCas9-chromatin modifier constructs in cytoplasmic (C) and nuclear extracts (N). Lysates were made from HEK293T cells that were transiently transfected with the dCas9-chromatin modifier constructs indicated. MEK1 and p53 were stained as cytoplasmic and nuclear controls respectively to determine success of fractionation. **B.** Western blot using a different set of cytoplasmic and nuclear lysates (derived as described above).

Notably, the dCas9-JMJD2A fusion protein does not appear to be present in the nuclear fraction, however there is a strong band at around 160 kDa in this fraction, which may correspond to a degradation product. It is possible that this construct is not imported into the nucleus or is present in amounts below the detection limit of the anti-Flag monoclonal antibody following fractionation. According to the manufacturer, the antibody is sensitive down to 10 ng protein, which for proteins around this size corresponds to $3x10^{10}$ molecules.

The dCas9-p300 construct is also not detectable on this blot (**Figure 3.3 A**),

however, it could be detected in the cytoplasmic and nuclear fraction in a separate experiment (**Figure 3.3 B**). In this second experiment, dCas9-TET1 was not detectable (likely due to a incomplete transfer to the blotting membrane).

3.1.2 Validation of constructs

Next, I sought to ascertain that the Cas9-chromatin modifier constructs are enzymatically active. Attempts at using immunoprecipitated protein in *in vitro* activity assays failed repeatedly. I successfully immunoprecipitated the dCas9-constructs as shown by Western blotting (**Figure 3.4**). However, the ELISA-based *in vitro* activity assays repeatedly showed low reproducibility between replicates, even with standards and control reactions included by the supplier of the assay kits. Thus, I eventually abandoned attempts to validate the constructs in this way.

While attempting the *in vitro* activity assays, a TALE-TET1 construct was reported [75] and I decided to test the dCas9-TET1 fusion construct I had made at one of the loci used in this publication. Both constructs have the exact same TET1 catalytic domain but differ only in the use of the DNA binding domain - TALE and dCas9 respectively. For validation of the dCas9-TET1 DNA demethylase construct I chose the *RHOXF2* locus used in the TALE-TET1 study. The RH3-TALE-TET1 construct from this study is used as a positive control as it has been shown to reliable remove methylation from CpG sites in the *RHOXF2* promoter. gRNA sequences that match RH3-TALE-TET1 binding site exactly could not be identified. The nearest gRNA target site that is followed by an NGG PAM sequence lies 5 bp upstream of the RH3-TALE-TET1 binding site and overlaps the TALE binding sequence (as indicated by arrows in **Figure 3.5**). When HeLa cells were transfected with the RH3-TALE-TET1 construct or the dCas9-TET1 construct plus gRNA, DNA demethylation relative to the GFP control could be detected at the locus by sequencing of bisulfite-converted DNA (21-24 clones analysed per sample). Levels of demethylation are similar for both constructs. The maximum observed decrease in methylation level is 30 % (see CpG number 8 in **Figure 3.5**), which is similar to changes reported previously with TALE-TET1 [75].

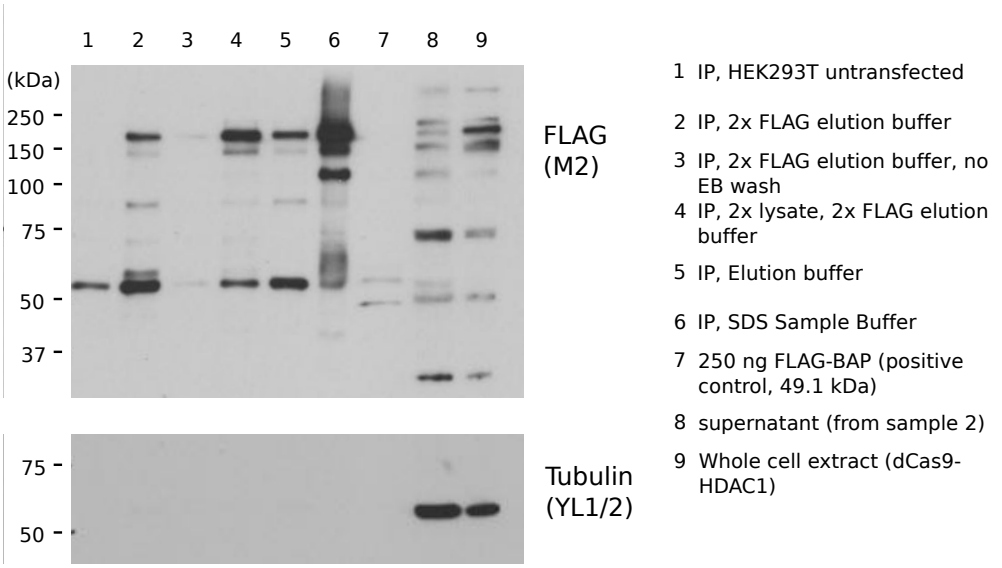


Figure 3.4: Immunoprecipitation of Cas9-chromatin modifier constructs from HEK293T cell lysates. Immunoprecipitation was carried out using the FLAG Immunoprecipitation Kit (Sigma) according to the manufacturer's instruction and immunoprecipitates were analysed by Western blotting using the Flag (M2) antibody. Lane1: Immunoprecipitate from untransfected HEK293T cells, Lane 2: Immunoprecipitate from HEK293T cells transfected with the dCas9-HDAC1 construct, protein was eluted using two elution steps with FLAG elution buffer, Lane 3: same as Lane 2 but with omission of a recommended wash step to remove unbound FLAG antibody from the resin, Lane 4: same as Lane 2 but using twice the recommended number of cells for lysis, Lane 5: same as Lane 2 but using Elution buffer (0.1 M Glycine, pH 3.5), Lane 6: same as Lane 2 but eluting in Sample buffer (125 mM Tris HCl, pH 6.8, with 4 % SDS, 20 % (v/v) glycerol, and 0.004 % bromophenol blue), Lane 7: Immunoprecipitate from untransfected HEK293T cells to which 250 ng FLAG-BAP protein was added (positive control), Line 8: Supernatant from Lane2, Line 9: Whole cell extract from HEK293T cells transfected with the dCas9-HDAC1 construct.

Given that DNA methylation yields digital information (either being absent or present), one might expect the level of methylation to be either 0 (unmethylated), 0.5 (hemi-methylated) or 1 (fully methylated) at any given CpG site. However, analysis of the GFP control reveals a mixed population. The fact that methylation does not decrease to 0 at the TET1 target site probably reflects the transfection efficiency. Not all cells have been successfully transfected. Selection was not applied in this experiment because the RH3-TALE-TET1 construct does not harbour a resistance gene for mammalian selection. In contrast, cells that have taken up the dCas9-TET1 construct could in principle be selected for with Hygromycin. In future, selection could be used to increase the percentage of cells expressing the construct and thus further increase the observed level of

demethylation.

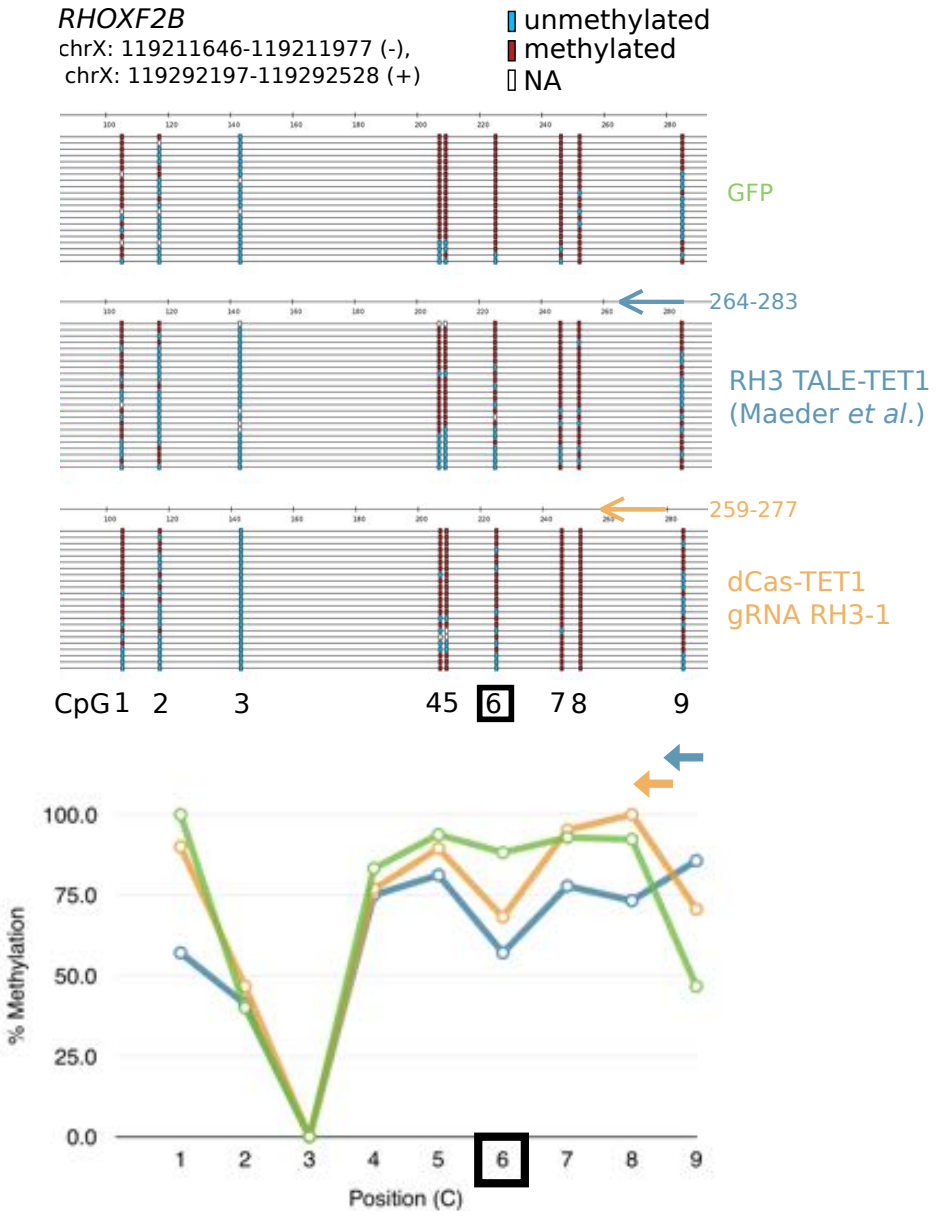


Figure 3.5: Validation of the dCas9-TET1 construct at the *RHOXF2* locus in HeLa cells. Cells were transfected with a GFP control (green), the RH3-TALE-TET1 construct from Maeder *et al.* [75] (blue), or the dCas9-TET1 wild-type construct + gRNA (orange). The constructs bind to their genomic target as indicated by the arrows. The dCas9-TET1 constructs bind slightly upstream of the TALE-TET1 construct. The target CpG is marked by a black box. Methylation level were quantified in each sample by sequencing of bisulfite-converted DNA (21-24 clones sequenced per sample) and are plotted for each of the 9 CpGs in the target amplicon.

In order to validate a given dCas9-chromatin modifier construct in a cell culture experiment rather than *in vitro* it is necessary to (1) identify a genomic locus

suitable for validation, (2) establish the baseline level of modification at the chosen locus, (3) design and synthesise gRNAs targeting the region and (4) assay dCas9-chromatin modifier activity at the chosen locus. I reasoned that performing validation for all remaining seven constructs in this way was not feasible in the time-frame of the PhD project. Several constructs similar to the ones I had produced have since become available [74, 78] and shown to work so I reasoned I could test the functionality of my constructs directly in the screening experiments.

3.1.3 Generation of stable cell lines expressing a dCas9 chromatin modifier

The dCas9-chromatin modifier sequences were cloned into the backbone of p-lenti-dCas9-VP64-Blast (Addgene plasmid 61425) for production of lentiviral particles. A dCas9-chromatin-modifier-T2A-Blasticidin fusion protein is expressed from these plasmids. The T2A peptide sequence is co-translationally cleaved [136]. In theory, this should ensure that cells only acquire Blasticidin resistance if the dCas9 chromatin modifier is successfully expressed.

Lentivirus was produced as described in the Methods section 2.1.7. A549 cells were transduced with lentivirus and polyclonal stable cell lines expressing one of the targetable chromatin modifiers were established. As expected and shown by immunofluorescence (IF, see **Figure 3.6**), expression levels in this population were heterogeneous. For screening, all cells should ideally express the same level of the dCas9-chromatin modifier to avoid attributing effects to targeting a particular site with a gRNAs that are in fact due to different levels of over-expression of the chromatin modifier.

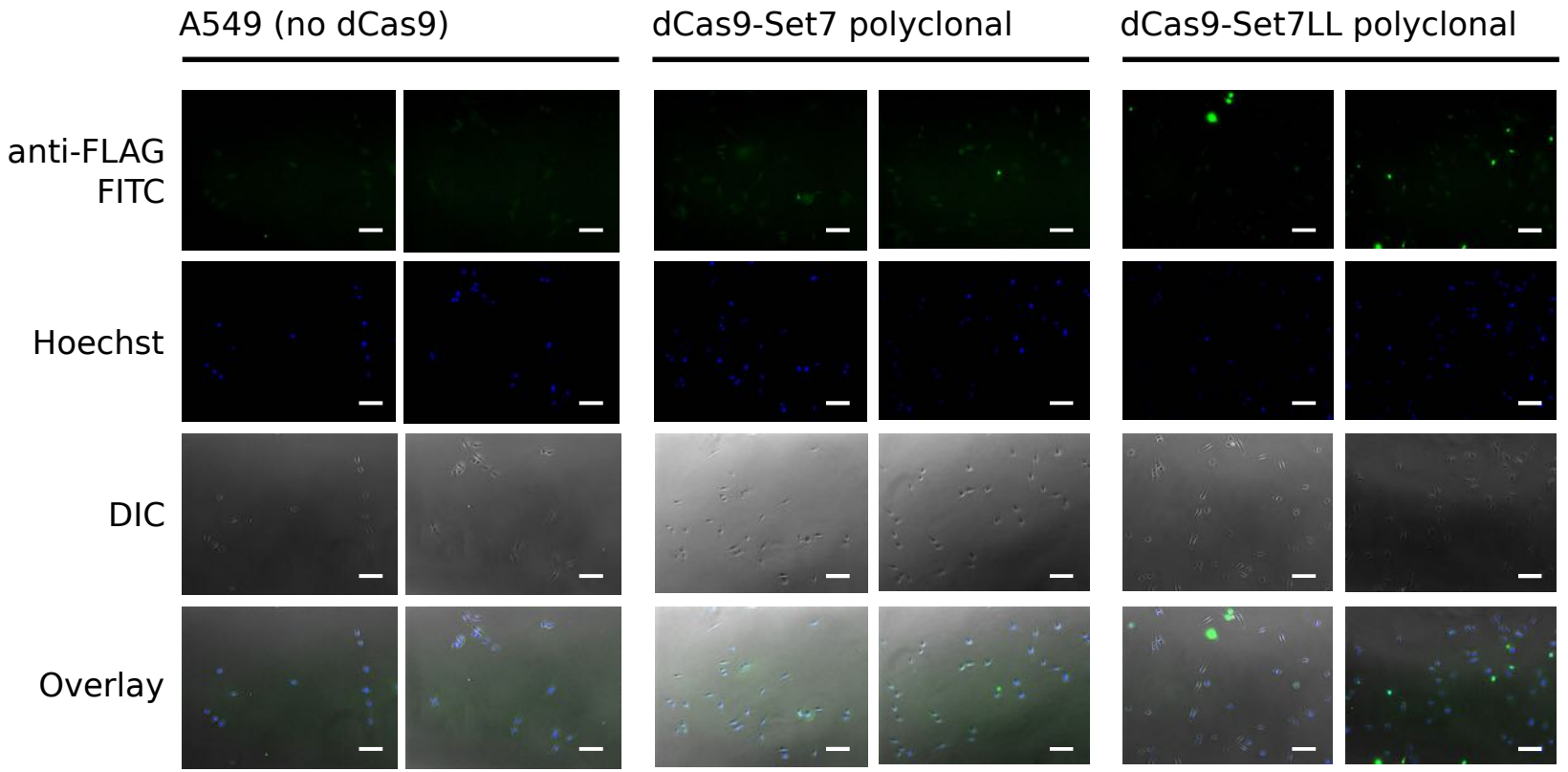


Figure 3.6: Representative example of immunofluorescence staining of polyclonal stable cell lines. Untransduced A549 cells or cells transduced with one of two lentiviral dCas9-SET7 constructs (LL indicates a lengthened linker in between dCas9 and SET7, this construct was kindly provided by Stefan Stricker) were fixed and stained using the anti-FLAG M2 FITC-conjugated antibody (Scalebar: 100 μm).

From the polyclonal population, monoclonal stable cell lines were established by serial dilution down to single-cell levels. Single-cell derived clonal populations were established and IHC was performed again to screen for clones that expressed dCas9 strongly. However, expression levels were much weaker compared to the brightly staining cells in **Figure 3.6** (data not shown). It could be that high expression levels of the dCas9-chromatin modifier fusions were not sustainable and the brightly staining cells simply died after serial dilution during re-growth. All monoclonal cell lines were clearly resistant against Blasticidin at the concentration used as untransduced control cells die within three days at the same level of selection. Based on showing signal slightly above background in IF staining, the following clonal lines were selected for further analysis: p300-C12, p300-C19, p300-C20, SET7-C1, SET7-C3, SET7-C4. Of these only the monoclonal cell line SET7-C3 expressed the dCas9-chromatin modifier at levels that were convincingly detectable by Western blot using the Flag M2 antibody (**Figure 3.7**). Please refer to section 3.3.1 for a discussion of these results and the rationale for nevertheless testing these cell lines in screening experiments.

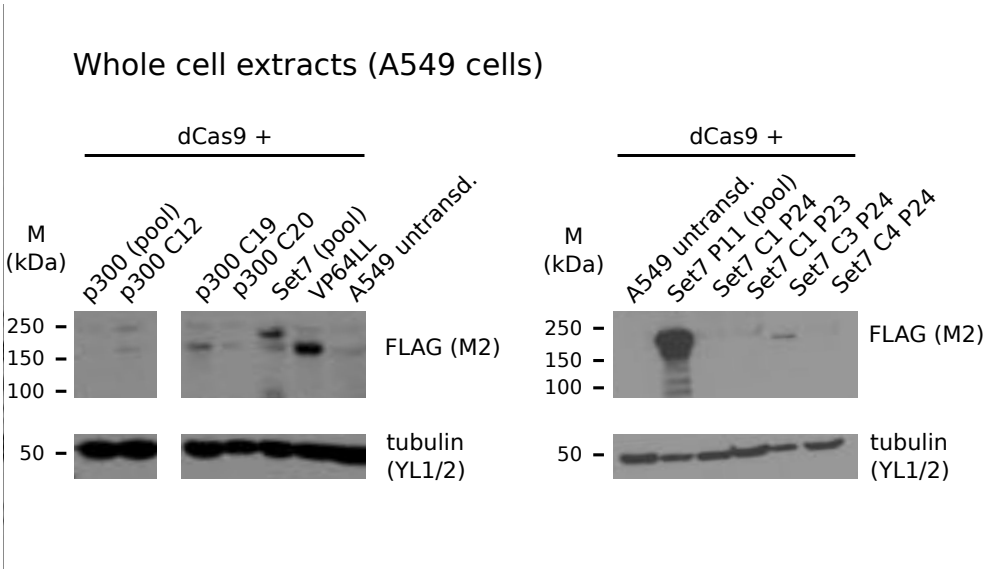


Figure 3.7: Western blot from whole-cell extracts of monoclonal stable cell lines (A549) expressing the dCas9-chromatin modifier indicated. Extracts from polyclonal cell lines (labelled as "pool") are loaded for comparison. Tubulin staining was used as a loading control.

3.2 Design of gRNA libraries

A gRNA library is a pooled collection of plasmids from which different gRNA sequences are expressed. For construction of gRNA libraries, I first modified the pLKO.1 plasmid backbone, commonly used for shRNA expression, to include a gRNA U6 promoter, 5'G from which transcription is initiated, the gRNA stem loop and terminator sequences instead of an shRNA expression cassette. This plasmid is named gRNA-pLKO.1 (**Figure 3.8**). Different gRNA targeting sequences can be cloned into the gRNA-pLKO.1 backbone downstream of the U6 promoter and upstream of the stem loop sequence to yield the complete library. Once introduced into cells the gRNA, consisting of targeting sequence and stem loop, will be transcribed. Cloning of gRNA sequences into the library vector is achieved through running circle amplification when the library of targeting sequences is constructed from a degenerate oligonucleotide that is used as a forward primer in the amplification reaction. When targeting sequences are generated as an oligonucleotide pool or from digested genomic DNA (see below) the mode of cloning is via a Gibson reaction, exploiting homology between the gRNA library plasmid gRNA-pLKO.1 and the flanking regions attached upstream and downstream of the targeting sequence [119].

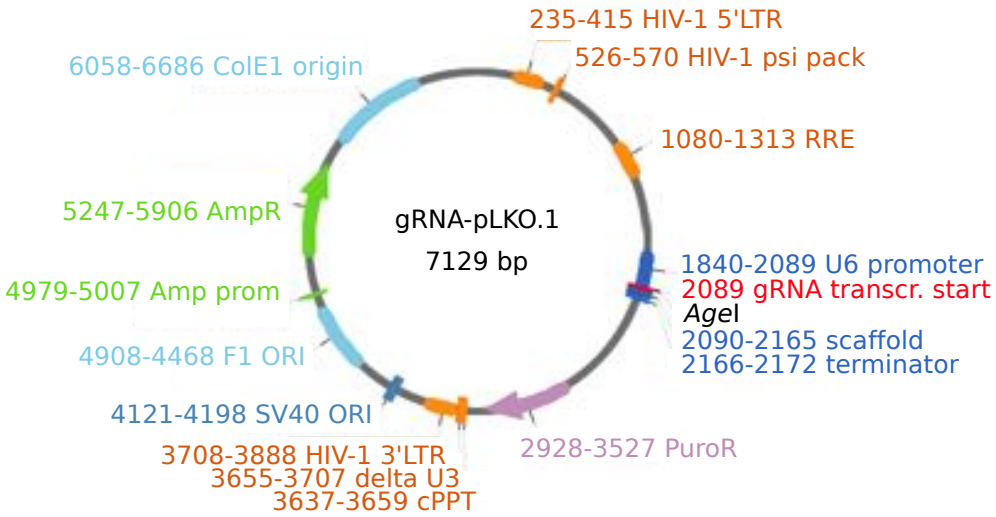


Figure 3.8: Map of the lentiviral gRNA expression plasmid gRNA-pLKO.1

3.2.1 Design and construction of a gRNA library targeting the promoters of 15 genes involved in EMT

To establish the epigenetic screening method using Cadherin switching as a read-out I designed a small-scale gRNA library targeting 15 genes that were reported to induce EMT when over-expressed in cells [123]. *In silico* design involved finding all gRNA sequences with an NGG PAM sequence in the human genome by searching for the string "$GN_{19}NGG$". The list of all possible locations was then intersected with the coordinates of regions of interest, here the promoter regions of the 15 genes identified as involved in regulation of EMT. The regions of interest where defined by visual inspection of the promoters of these genes in UCSC Genome Browser [137] and selection of regions where chromatin modifications are located around the promoter (see **Table 2.9** in the Methods chapter). This yielded a total of 5,086 gRNA sequences (**Figure 3.9**). These were ordered as a custom oligonucleotide pool from Custom Array and cloned into the library backbone by Gibson assembly as described in the Methods section 2.2.7. The resulting library is referred to as the **EMT5000 library.**

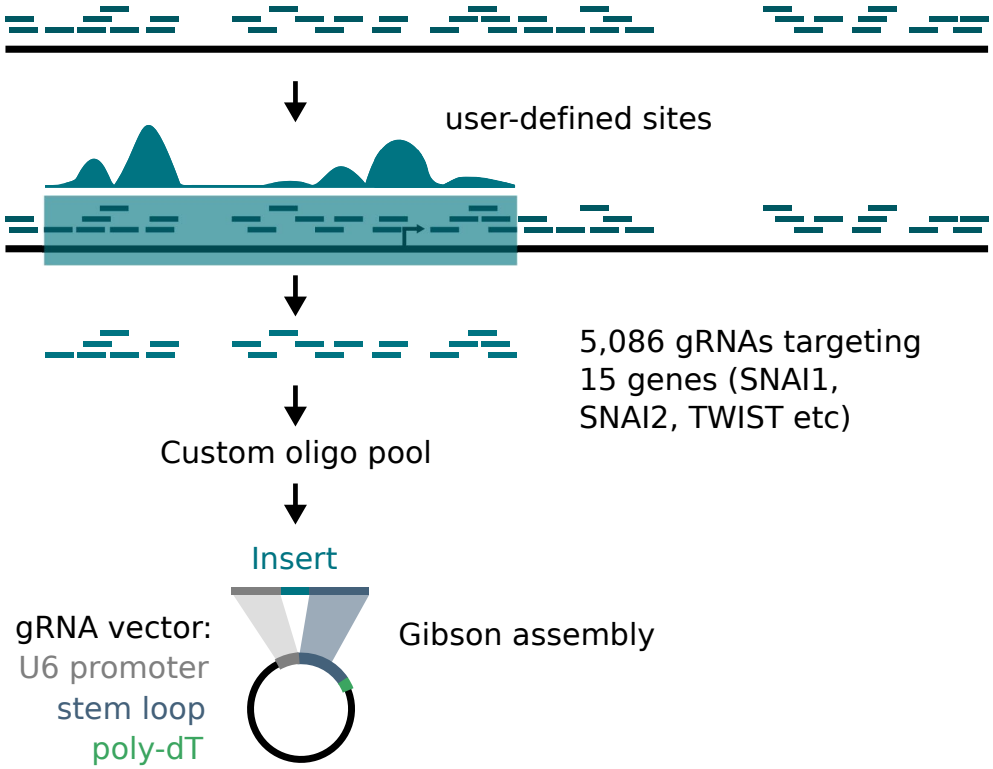


Figure 3.9: *In silico* design and construction of the EMT5000 library. All gRNA sequences of the form $GN_{19}NGG$ are found in the human genome. These are intersected with user-defined regions of interest, here corresponding to the promoter regions of 15 genes involved in regulation of epithelial-to-mesenchymal transition. The resulting 5,086 sequences are ordered as a custom oligonucleotide pool and inserted into the gRNA library backbone by Gibson cloning.

3.2.2 Generation of ultra-complex gRNA libraries

Because chip-based custom oligonucleotide pool synthesis is costly and currently limited to around 100,000 sequences per chip, I wanted to explore different ways of generating ultra-complex gRNA libraries. These could in future be useful for genome-wide screens, especially when the aim is to target non-coding regions which are much broader than exons and transcriptional start sites, for which gRNA libraries have already been generated (**Figure 1.7**). I considered two possible strategies: (1) Design of a degenerate oligonucleotide sequence enriched for sequences that target regulatory regions such as promoters and (2) enzymatic digestion of genomic DNA down to fragments roughly 20 bp in size and cloning of these fragments into the gRNA vector.

Bioinformatic analysis was used to identify degenerate oligonucleotide sequences that enrich for binding to promoters (see Methods section 2.2.1 for details), which

were then used for the construction of the **degenerate gRNA libraries**. First, all gRNAs that fall into putative promoter regions were identified in the repeat-masked human genome. This yielded 4,113,530 gRNA sequences. In an effort to reduce complexity, sequences were clustered based on sequence identity and a consensus oligonucleotide sequence that would represent each cluster was derived. This step improved targeting efficiency (i.e. the number of promoter hits relative to the number of sequences in the library) by an order of magnitude compared to a random consensus sequence of the same complexity. The library is made by synthesising a degenerate, phosphorylated oligonucleotide and using it as a primer in circle amplification. Intramolecular ligation of the PCR product incorporates the gRNA targeting sequence into the vector. The resulting libraries are much more complex and much more cost-effective than a synthetic library like the EMT5000 library. However, due to the need to represent all the desired target sequences by a single, degenerate sequence, some of the sequences encoded in the consensus sequence do not align to the genome at all, and some desired targeting sequences are lost, hence an overall large decrease in targeting efficiency compared to the EMT5000 library is to be expected for this approach.

A **random library** was also produced by ordering a random GN_{19}-oligonucleotide and using it in circle amplification. This library is very complex, should be genome-wide and should theoretically cover all gRNA target sites in the genome (although this depends on the scale of synthesis). Although extremely cost-effective to produce, gRNA libraries based on degenerate oligonucleotides are probably of limited use in practice. Too many gRNAs in these libraries do not map to the genome at all or do not fall into regions of interest, which decreases overall targeting efficiency (see **Table 3.1**).

I next turned to enzymatic methods to generate gRNA libraries from an existing source of DNA rather than using *de-novo* oligonucleotide synthesis. I reasoned that it should be possible to use a relatively non-specific nuclease to digest genomic DNA into pieces roughly 20 bp in size. While digestion with DNase proved to be poorly controllable, digestion with micrococcal nuclease (MNase) could be controlled by varying the amount of enzyme (**Figure 3.10 B**). A strategy to derive a gRNA library from fragments of genomic DNA generated by micrococcal nuclease digestion is shown in **Figure 3.10 A**. The optimal amount of MNase to digest genomic DNA down to 15-50 bp in size was determined and fragments of the desired size were excised from a PAGE gel (**Figure 3.10 C**). Following end repair, adapter sequences were ligated to the genomic DNA fragments. Adapters are amplified from the gRNA-pLKO.1 backbone and digested with a restriction

Name	**Consensus sequence**	**Complexity** (num. gRNAs)	**Promoter hits** (no MM)	**Promoter hits** (3 MM)	**Unique promoters hit** (0 MM)	**Unique promoters hit** (3 MM)	**Targeting efficiency** (Promoter Hits/Complexity)
C190	GRRRGRGRRRRRGRRGRRGV	24,576	223	69,399	212	22,655	9.07E-03
C304	GGGKKKGKGKKGKKKGKKBS	12,288	75	22,055	72	9,583	6.1E-03
C422	GRRRRRRRRRRRRRRRRGR	262,144	1,556	231,923	1,221	39,734	5.94E-03
C369	GGRRRGRGRRRRRGRRRSVR	49,152	279	84,309	269	26,219	5.68E-03
C1281	GGRGRRRRRGRRRRRGSRRV	49,152	205	62,711	197	21,439	4.17E-03
C89	GVRVSSSSSGSSGSSRGSSK	147,456	137	75,791	134	20,389	9.29E-04
TOTAL	combine and uniquify	**510,933**	**2,141**	**371,017**	**1655**	**49,701**	**4.19E-03**
Random1	GNNNNNNNNNNNNNNNNNN	274,877,906,944	467,1728	-	82,903	-	1.7E-05
Random2	GSSSSSSSSSSSSSSSSSS	524,288	274	421,413	261	26,072	5.23E-04

Table 3.1: Characteristics of six consensus oligonucleotide sequences derived using a clustering algorithm to enrich for sequences targeted to promoter regions. The table includes the number of different sequences encoded by the degenerate oligonucleotide sequences ("complexity"), how many of these hits the consensus has in promoters allowing 1 mismatch (MM) or 3 mismatches anywhere in the targeting sequence, as well as how many promoters out of 91,433 putative promoter regions (-1000 to +500 bp around the start of all known transcripts, human GRCh37-70, downloaded from Ensembl) are hit with this consensus sequence. The targeting efficiency of each library is calculated by dividing the number of promoter hits by the library complexity and compared to those of a random GN_{19} library as well as a random consensus of the same complexity.

enzyme to ensure directionality of ligation. Adapter dimer formation should not occur because the adapters lack the 5' phosphate required for ligation as they are generated by PCR using unphosphorylated primers. 5'-phosphates are only present on successfully end-repaired genomic DNA fragments. Three different pairs of PCR primers against the adapter sequences were used to amplify the ligation product, yielding double-stranded DNA fragments suitable for use in Gibson assembly. Fragments onto which both the adapter bearing the U6 sequence and the adapter derived from the gRNA stem loop had been ligated were selected by gel excision. Different length amplicons (referred to as "L", "M" and "S") were used as input for the Gibson reaction to generate the **MNase digest library**.

Next-generation sequencing revealed that the MNase digest libraries made from human (**Figure 3.11** bottom centre) or mouse (**Figure 3.11** bottom right) genomic DNA are of extremely high complexity, with few sequences being sequenced more than once. Using a Bayesian model it was possible to estimate the complexity of the original library based on the sample of the population obtained by sequencing and the data are consistent with a library complexity of $5x10^7$-10^9 individual gRNA sequences (This Bayesian model was contributed by Dr. Christiane Fuchs, Helmholtz Institute Munich, Germany). It also became evident from the sequencing data that using long overhangs lead to a more efficient incorporation into the vector as there were fewer empty or truncated gRNA sequences in the human "L" vs "S" sample. Compared to the libraries made from synthetic oligonucleotides, the MNase libraries show a broad distribution of gRNA lengths (**Figure 3.12** bottom right and centre), ranging from 18 to 46 bp with a median length of 28 bp. The random and degenerate libraries, generated from a synthetic oligonucleotide have an expected size of 20 bp with only a small minority of reads harbouring shorter or longer gRNA protospacer sequences (**Figure 3.12** grey and blue panels). These either reflect errors introduced during sequencing, PCR, Gibson assembly or oligonucleotide synthesis.

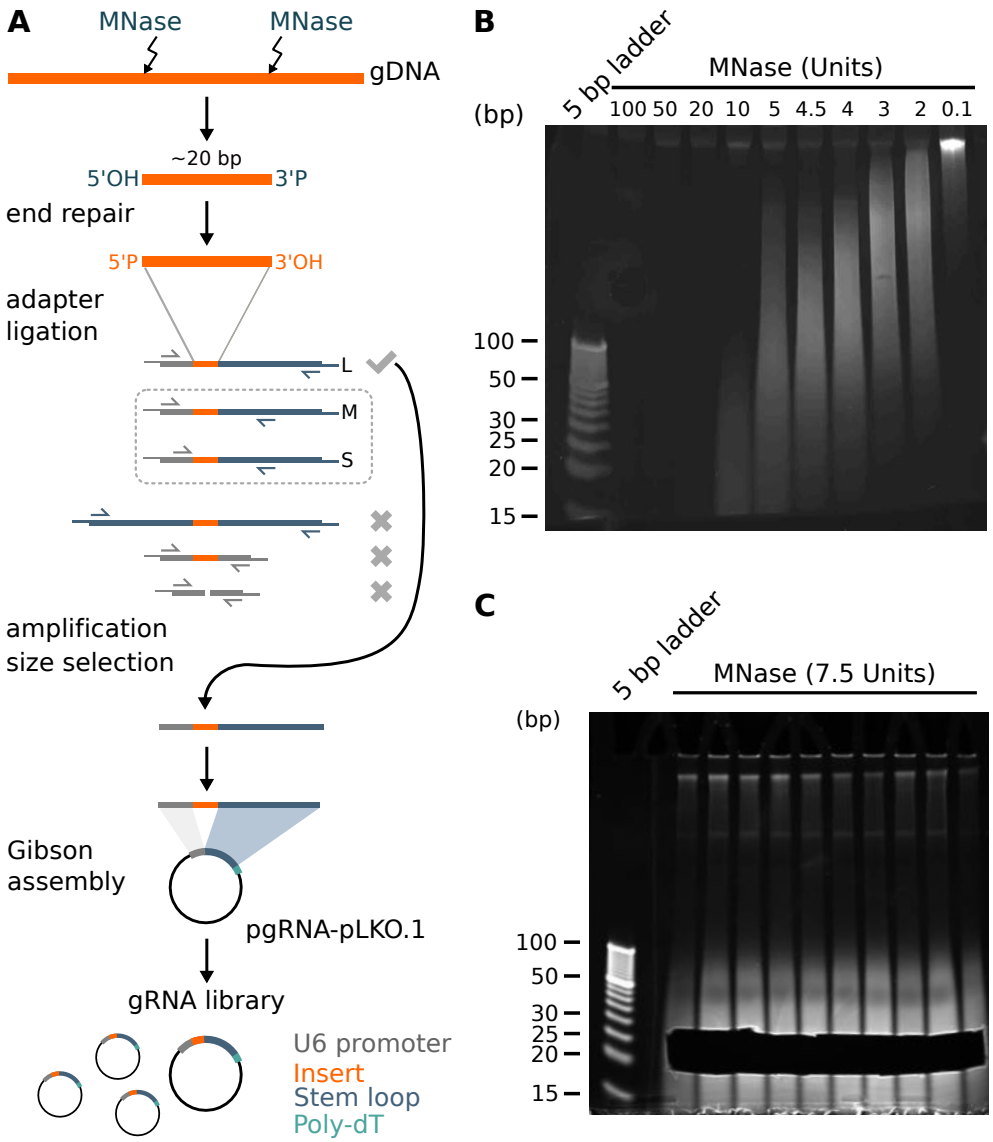


Figure 3.10: Construction of gRNA libraries from micrococcal nuclease-digested genomic DNA. A. Strategy. Purified genomic DNA is digested down to roughly 20 bp fragments with micrococcal nuclease (MNase). Following end repair, the genomic fragments are ligated to a pair of different-size adapters and amplified by limited-cycle PCR. Following size selection to isolate only fragments with the correct adapters ligated to either end, the fragments are inserted into the gRNA vector using Gibson cloning to yield a library of targeting sequences. **B.** Genomic DNA digested with various amounts of MNase are analysed on a 20 % PAGE gel. **C.** Fragments in the range of 17-35 bp range were cut from the gel and isolated for library construction.

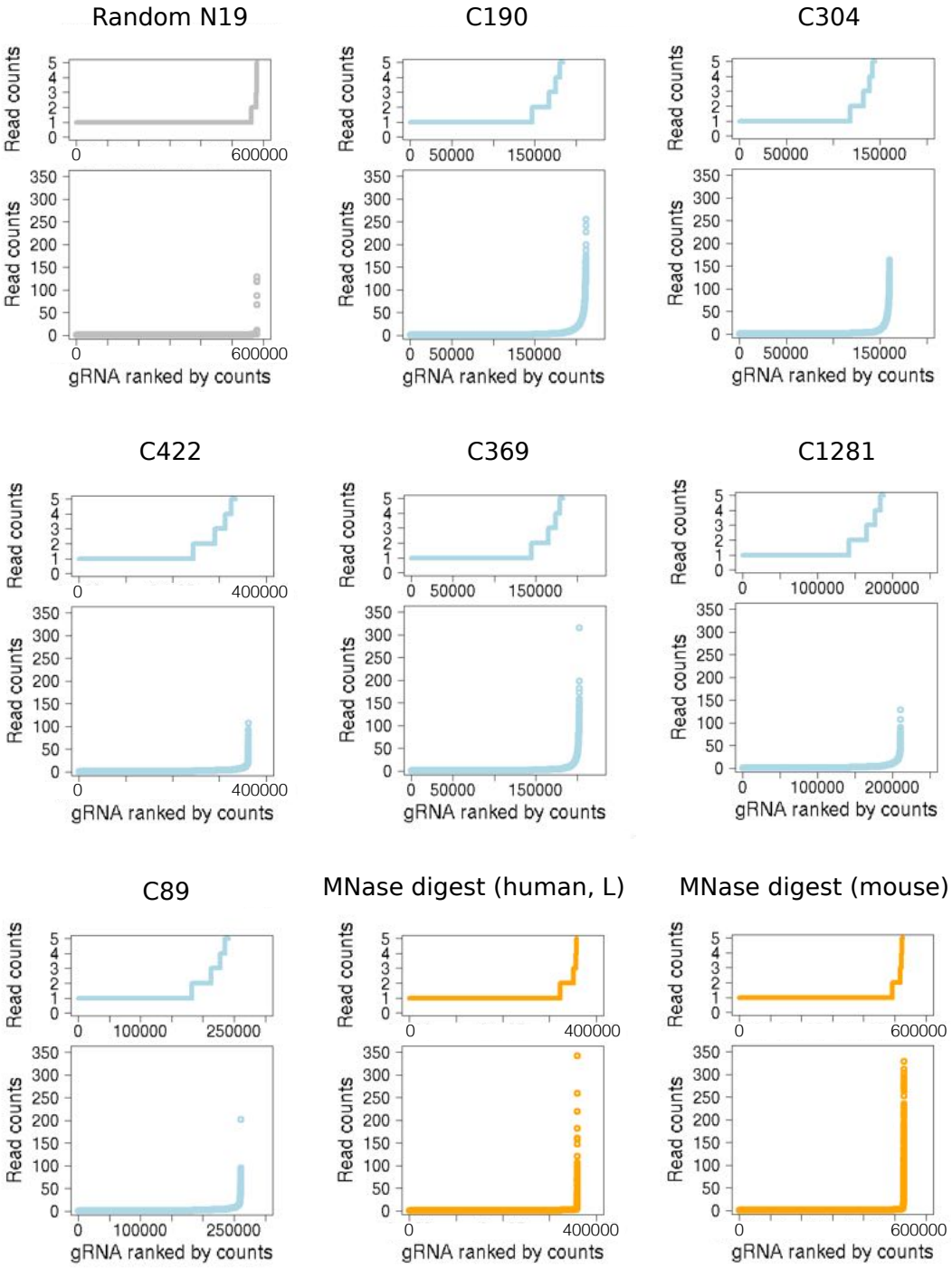


Figure 3.11: Library QC by sequencing - Plots of read counts for each sequenced gRNA. For each library, top graphs depict the area between read counts 0 and 5 from the lower graphs. The random library is displayed in grey. The degenerate libraries are in light-blue. The MNase digest libraries, made from enzymatic digestion of human or mouse genomic DNA, are shown in orange. Most gRNA sequences are sequenced once suggesting large sequence diversity.

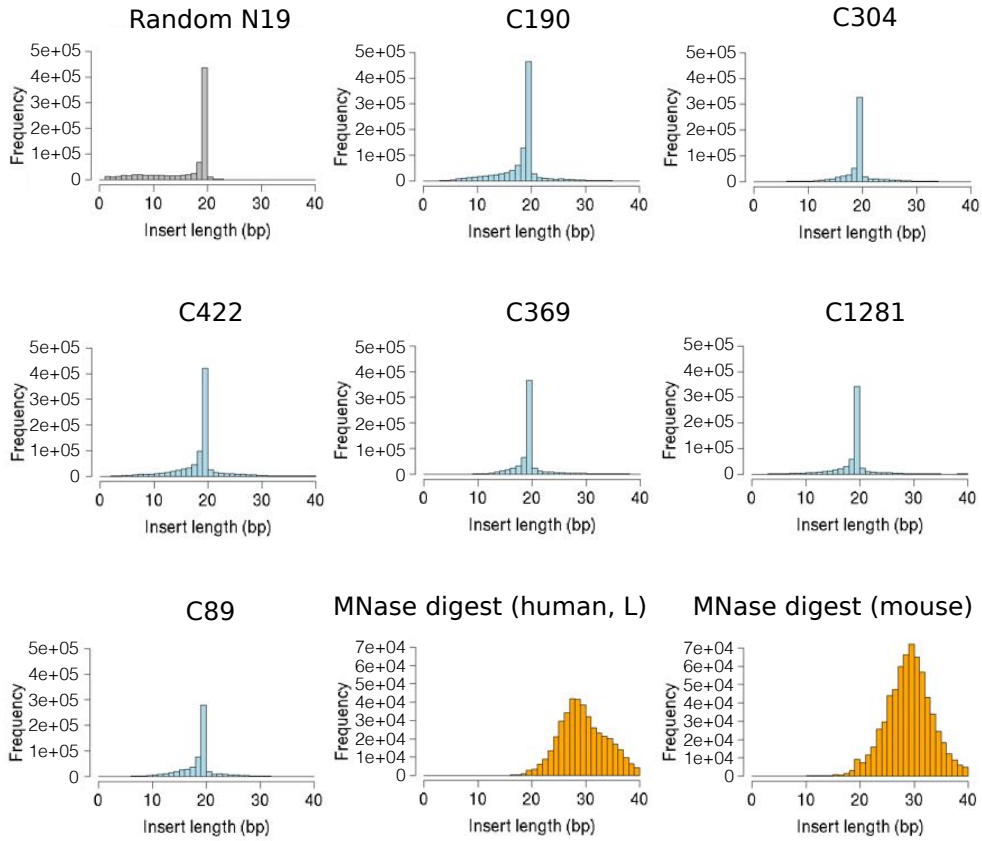


Figure 3.12: Library QC by sequencing - Distribution of gRNA read lengths for each of the libraries. Read counts (frequency) is plotted against the length of the gRNA targeting sequence that was determined by sequencing. The random library (grey) and degenerate libraries (light-blue), were made from circle amplification using a synthetic oligonucleotide primer. The MNase digest libraries (orange) were made from enzymatic digestion and cloning of human or mouse genomic DNA and show a broader read length distribution.

While a length of 20 bp is generally desired for a sequence to work as a gRNA, it has recently been shown that longer 30 bp gRNAs do work and that they appear to be trimmed down to 20 bp from the 5' end by the cellular RNA processing machinery [138]. Given that a sizeable fraction of gRNAs in the MNase digest libraries are even longer than 30 bp, I wondered whether these would also still be functional. I randomly selected gRNA protospacer sequences of length 35, 40 and around 45 bp that mapped uniquely to the human genome with an NGG PAM from the sequencing data generated from the human MNase digest library. I cloned these individual gRNAs into the vector pX458, which harbours both a gRNA expression cassette as well as a wild-type Cas9-GFP fusion sequence. HEK293T cells were transiently transfected with these constructs and transfection efficiency was found to be around 30 % based on GFP expression (see **Figure 3.13 B** for a representative example). Cas9 introduces a double-strand

break 3 bp upstream of the PAM sequence. In the absence of a repair template, the targeted lesion is repaired by non-homologous end joining (NHEJ), which is error-prone and often leads to the generation of small insertion/deletion (indel) mutations. Such Cas9-induced mutations at the gRNA target site are detectable by next-generation sequencing of PCR amplicons (**Figures 3.13, 3.14, 3.15, 3.16, 3.17**). This shows that even gRNAs 44 and 46 bp in length (**Figure 3.13** and **Figure 3.14**) can still direct Cas9 to its genomic target site. For all five gRNAs tested here the shortened 20 bp versions appeared to be slightly more efficient (**Figures 3.13 C, 3.14 B, 3.15 B, 3.16 B, 3.17 B**) at inducing targeted indels together with wild-type Cas9 compared to the longer gRNAs.

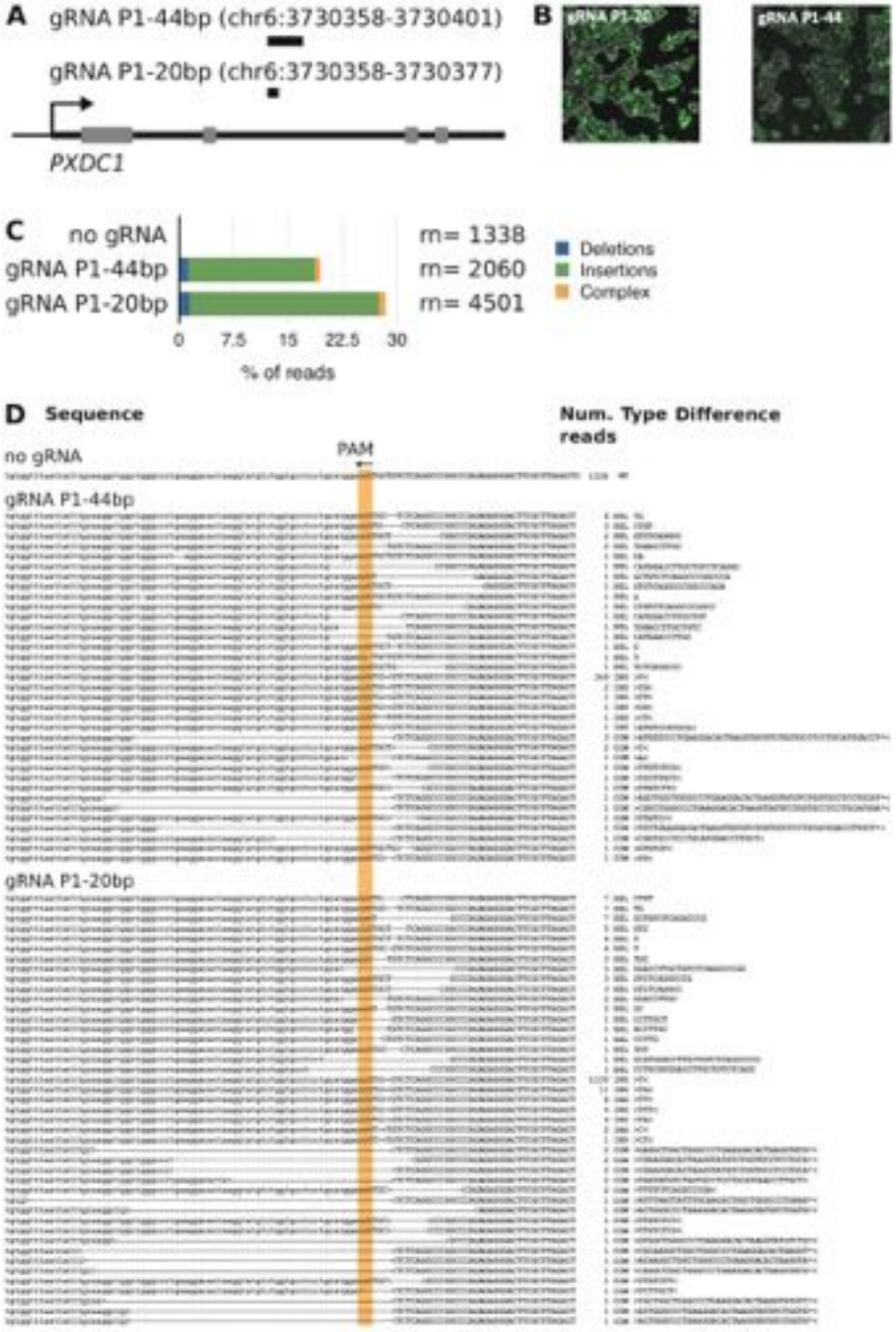


Figure 3.13: Validation of gRNAs longer than 30 bp from MNase library: P1-44.

Figure 3.13: Validation of gRNAs longer than 30 bp from MNase library: P1-44 (continued). gRNA P1-44 bp was selected from the sequencing data of the human MNase digest library. **A.** Diagram showing target sites of the gRNA and of its shortened 20 bp version. **B.** A wild-type Cas9-GFP fusion protein is also expressed from the vector used for gRNA expression (px458). Microscopy reveals around 30 % transfection efficiency based on the number of GFP-positive cells. **C.** Percentages of reads harbouring mutations in next-generation sequencing of the gRNA target sites in untransfected HEK293T cells and those transfected with the long and short versions of the gRNA. **D.** List of the most common mutations in the sequencing data from the different samples together with the number of reads and type of mutation: Ins (insertion), Del (deletions), Com (Complex indel). The difference to the wild-type sequence is also indicated showing sequence that was lost or inserted respectively. The gRNA protospacer is marked by capital letters and the PAM sequence is highlighted in orange.

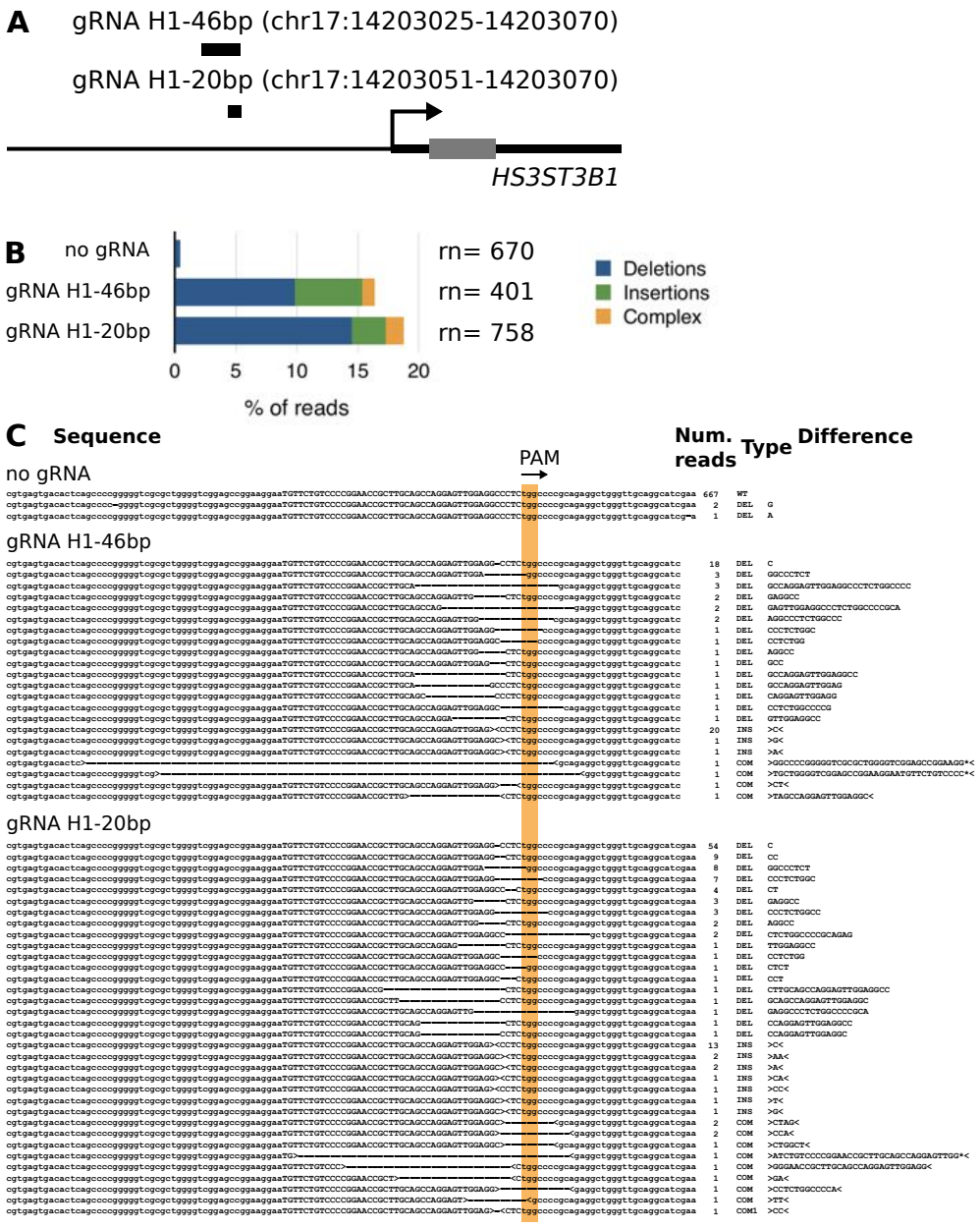


Figure 3.14: Validation of gRNAs longer than 30 bp from MNase library: H1-46. gRNA H1-46 bp was selected from the sequencing data of the human MNase digest library. **A.** Diagram showing target sites of the gRNA and of its shortened 20 bp version. **B.** Percentages of reads harbouring mutations in next-generation sequencing of the gRNA target sites in untransfected HEK293T cells and those transfected with the long and short versions of the gRNA. **C.** List of the most common mutations in the sequencing data from the different samples together with the number of reads and type of mutation: Ins (insertion), Del (deletions), Com (Complex indel). The difference to the wild-type sequence is also indicated showing sequence that was lost or inserted respectively. The gRNA protospacer is marked by capital letters and the PAM sequence is highlighted in orange.

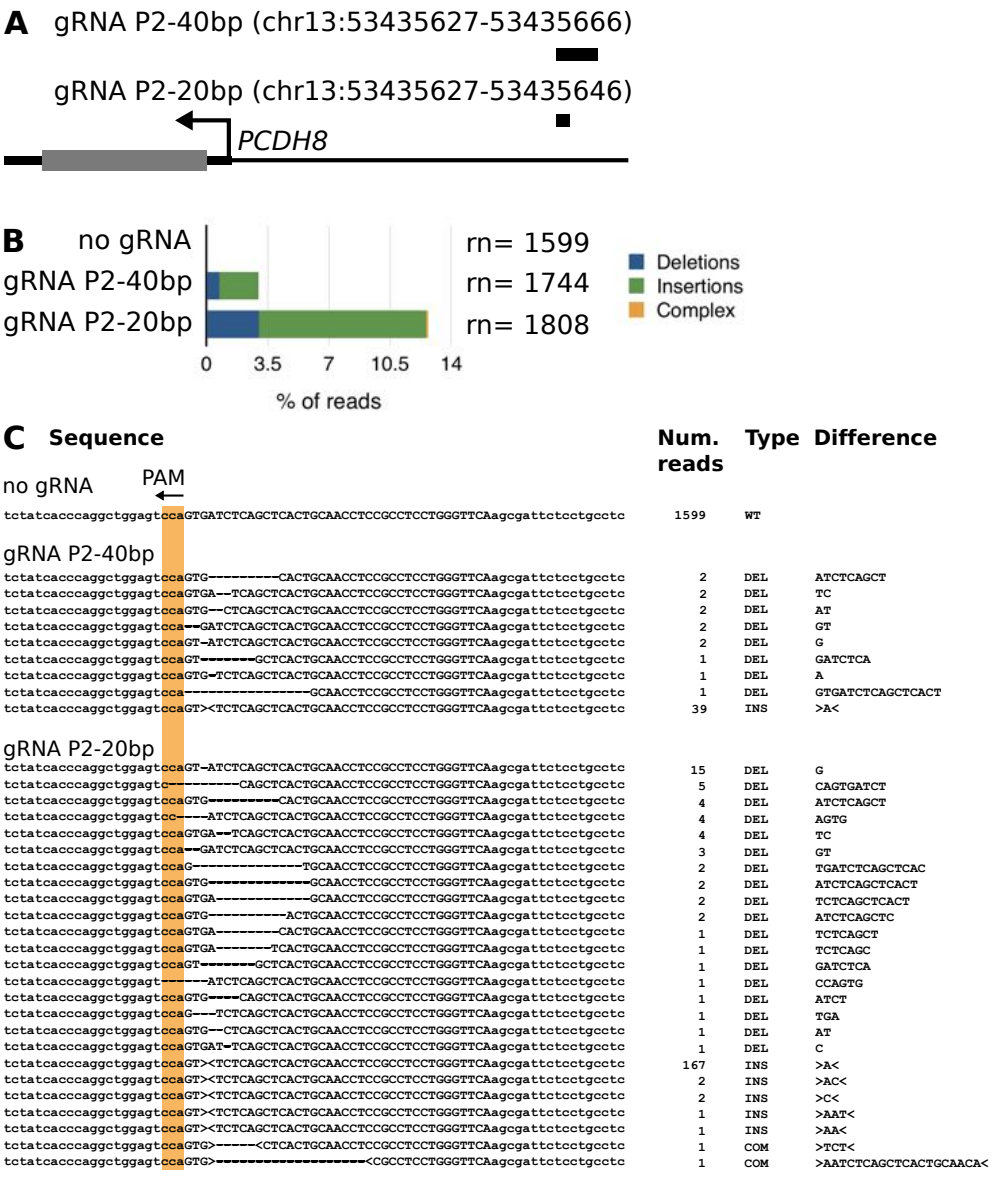


Figure 3.15: Validation of gRNAs longer than 30 bp from MNase library: P2-40. gRNA P2-40 bp was selected from the sequencing data of the human MNase digest library. **A.** Diagram showing target sites of the gRNA and of its shortened 20 bp version. **B.** Percentages of reads harbouring mutations in next-generation sequencing of the gRNA target sites in untransfected HEK293T cells and those transfected with the long and short versions of the gRNA. **C.** List of the most common mutations in the sequencing data from the different samples together with the number of reads and type of mutation: Ins (insertion), Del (deletions), Com (Complex indel). The difference to the wild-type sequence is also indicated showing sequence that was lost or inserted respectively. The gRNA protospacer is marked by capital letters and the PAM sequence is highlighted in orange.

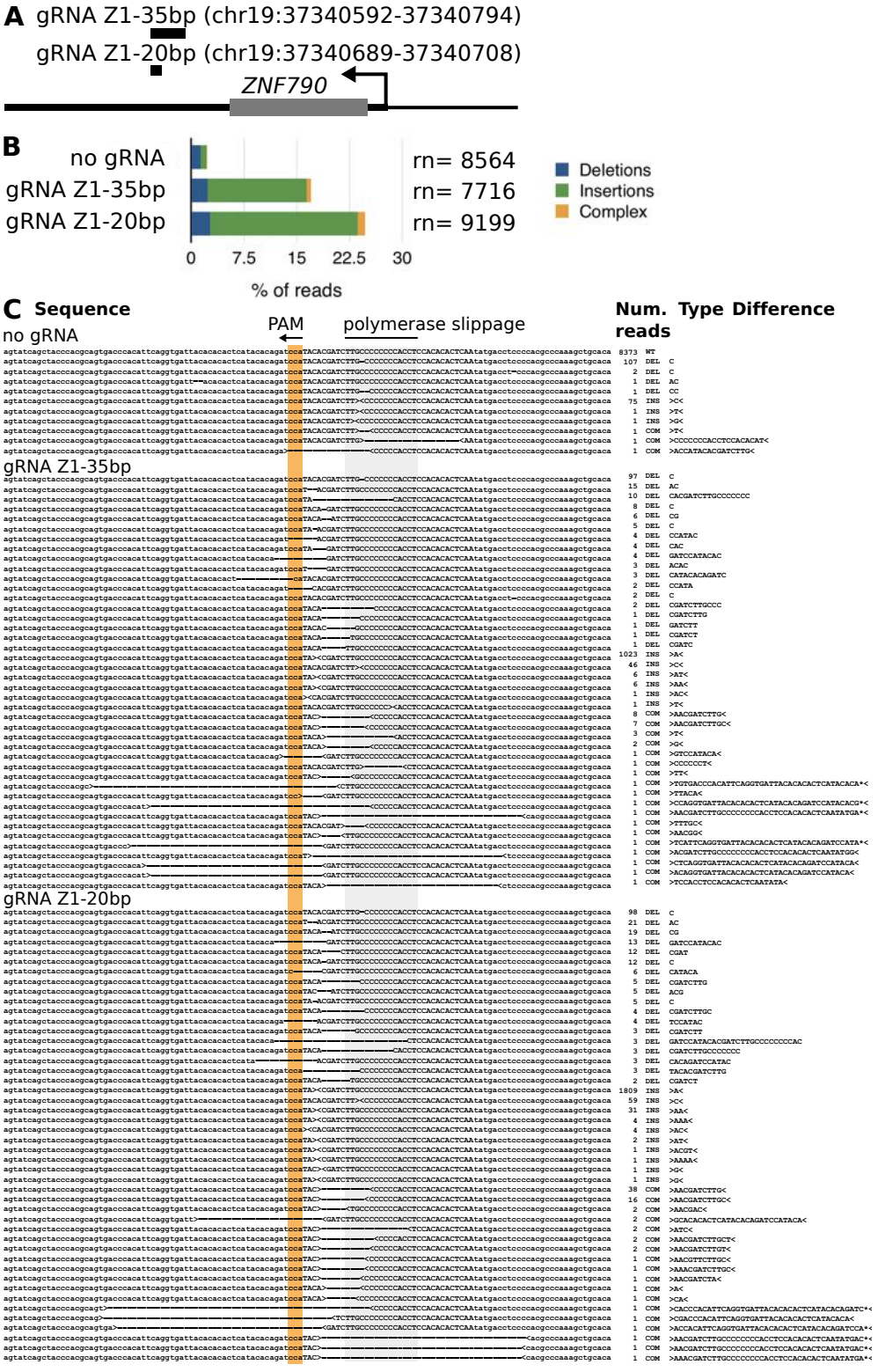


Figure 3.16: Validation of gRNAs longer than 30 bp from MNase library: Z1-35.

Figure 3.16: Validation of gRNAs longer than 30 bp from MNase library: Z1-35 (continued). gRNA Z1-35 bp was selected from the sequencing data of the human MNase digest library. **A.** Diagram showing target sites of the gRNA and of its shortened 20 bp version. **B.** Percentages of reads harbouring mutations in next-generation sequencing of the gRNA target sites in untransfected HEK293T cells and those transfected with the long and short versions of the gRNA. **C.** List of the most common mutations in the sequencing data from the different samples together with the number of reads and type of mutation: Ins (insertion), Del (deletions), Com (Complex indel). The difference to the wild-type sequence is also indicated showing sequence that was lost or inserted respectively. The gRNA protospacer is marked by capital letters and the PAM sequence is highlighted in orange. Mutations in the untransfected control sample resulting from polymerase slippage are highlighted in grey.

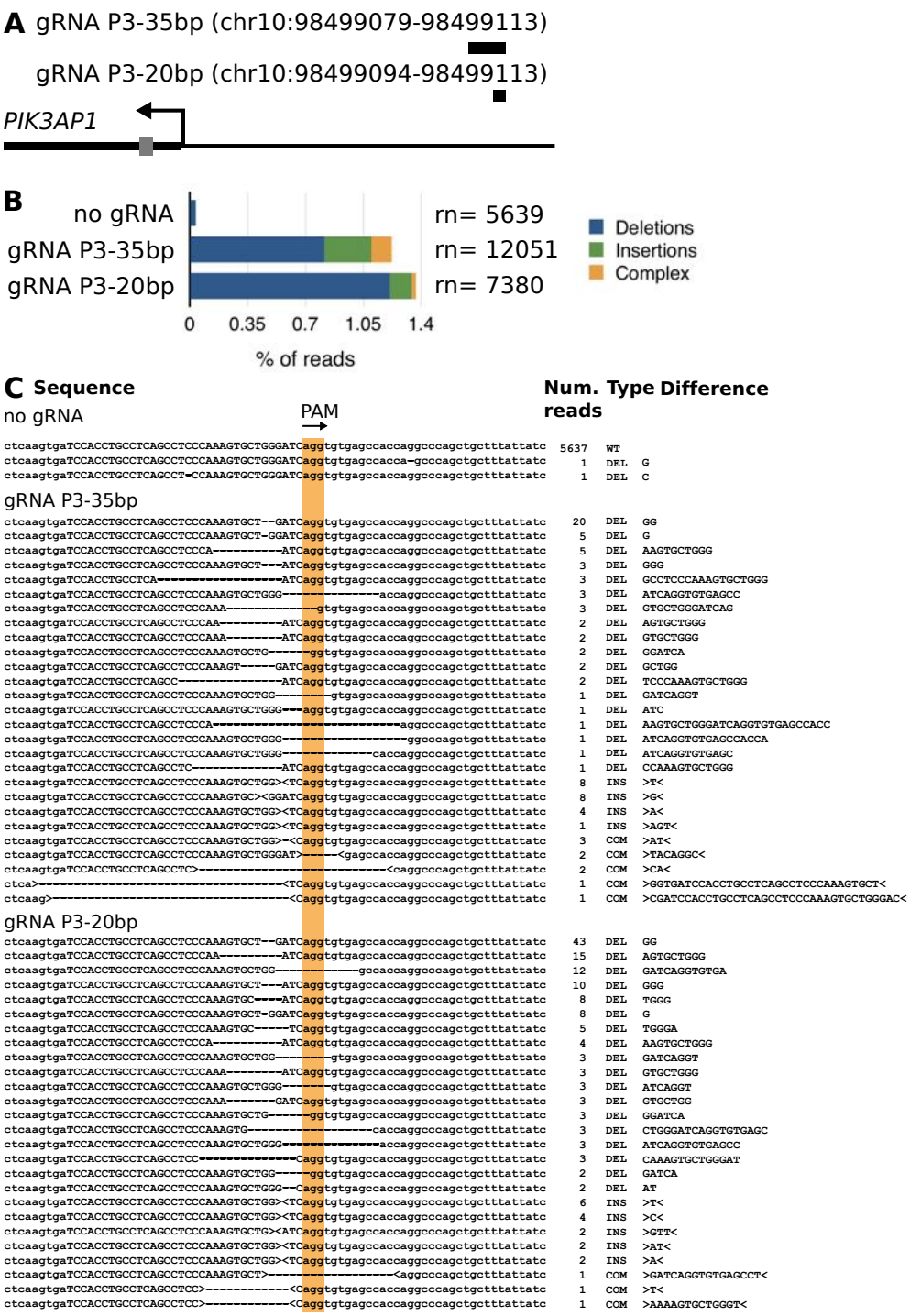


Figure 3.17: Validation of gRNAs longer than 30 bp from MNase library: P3-35.

Figure 3.17: Validation of gRNAs longer than 30 bp from MNase library: P3-35 (continued). gRNA P3-35 bp was selected from the sequencing data of the human MNase digest library. **A.** Diagram showing target sites of the gRNA and of its shortened 20 bp version. **B.** Percentages of reads harbouring mutations in next-generation sequencing of the gRNA target sites in untransfected HEK293T cells and those transfected with the long and short versions of the gRNA. **C.** List of the most common mutations in the sequencing data from the different samples together with the number of reads and type of mutation: Ins (insertion), Del (deletions), Com (Complex indel). The difference to the wild-type sequence is also indicated showing sequence that was lost or inserted respectively. The gRNA protospacer is marked by capital letters and the PAM sequence is highlighted in orange.

Additional characterisation of the MNase digest libraries based on the sequencing data I generated was performed by my collaborator Karolina Worf at the Helmholtz Centre in Munich. The results of this analysis are depicted in **Figure 3.18**. The protospacer sequences that align to a ribosomal gene unit on chromosome 21 in the three human MNase digest libraries that were sequenced separately are shown in **Figure 3.18 A**. The gRNAs incorporated into the MNase digest libraries appear to be evenly distributed across the human and mouse genomes respectively (**Figure 3.18 C**). As expected, given that I have not found a way to ensure the presence of NGG PAM sequences downstream of the gRNA protospacer, many gRNAs in the library do not map to sites that are followed by a particular PAM. 25 % of gRNAs harbour an *S.pyogenes* PAM immediately downstream of the target site (**Figure 3.18 B**). An alternative method for the generation of large-scale gRNA libraries was recently reported. This method, called "CRISPR-Eating" [122], uses the restriction enzymes *Hpa*II, *ScrF*I, and *Bfa*I, which recognise the sequences C/CGG, CC/NGG, and C/TAG respectively, to generate the 3' end of the gRNA. This should ensure that an *S.pyogenes* PAM is present at the 3' end of the gRNA, however it introduces the additional restriction that all gRNAs have a C at the 3' end. Thus, not all possible gRNAs are accessible using the CRISPR-Eating method, compared to the MNase digestion protocol (**Figure 3.18 D**). Sequencing of the CRISPR-Eating library generated for the *E.coli* genome revealed that desired 20 bp gRNA protospacers that are followed by an *S.pyogenes* PAM accounted for 44 % of the total material sequenced. "Of the remaining 56 %, 45 % of the total material consisted of guides shifted by one to three bases 3' relative to PAMs, likely due to promiscuous activity of Mung-bean nuclease used to blunt fragments", according to Lane *et al.* [122]. Therefore, while elegant on paper, the CRISPR-Eating method does not always ensure the presence of a PAM sequence 3' of the cloned protospacer sequence.

Comparison of the GC content of the sequenced libraries relative to the genomes they were generated from shows that all three methods introduce a slight GC bias, possibly introduced by PCR amplification, with the human MNase digest libraries showing a slightly increased bias (**Figure 3.18 E**). Because the human and mouse MNase digest libraries were generated in parallel, this additional GC bias is unlikely to be the result of PCR amplification bias. It could be that the bias was already present in the original sample of human gDNA extracted from Peripheral Blood Mononuclear Cells (PBMCs) from pooled blood (generous gift of Lee M. Butcher). The mouse genomic DNA was from a commercial source.

Although considered a relatively non-specific endonuclease, a preference for cutting at dXp-dTp and dXp-dAp bonds has been reported for MNase [139]. Examination of the nucleotide composition around the cut sites revealed an A/T enrichment at the first and last base of the protospacer for the MNase digest libraries. As expected given the recognition sites of *Hpa*II, *ScrF*I, and *Bfa*I, the last and +1 position in the CRISPR-Eating libraries are almost exclusively occupied by cytosines.

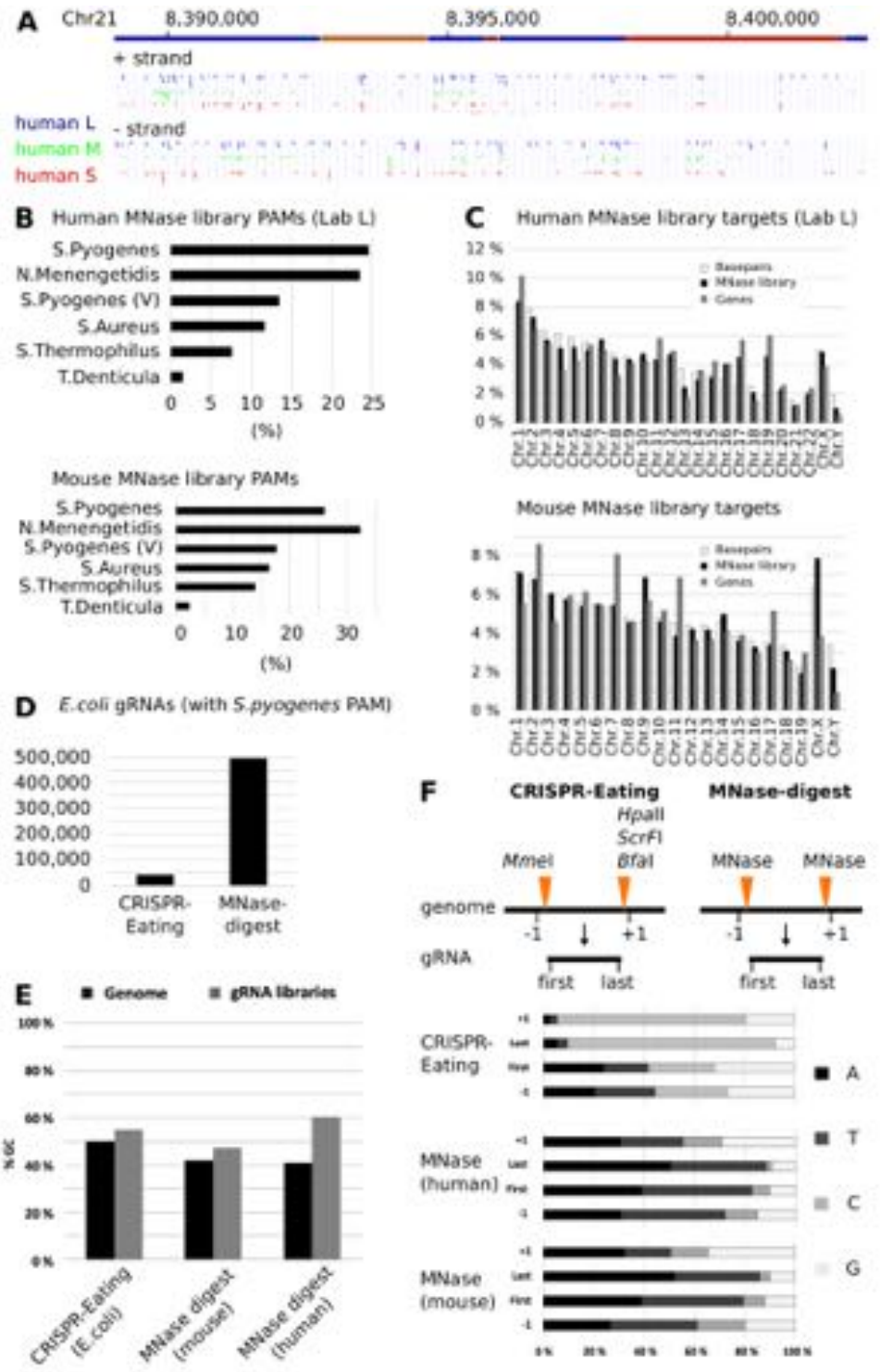


Figure 3.18: Characteristics of the MNase digest libraries and comparison to CRISPR-EATING method.

Figure 3.18: Characteristics of the MNase digest libraries and comparison to CRISPR-EATING method (continued). The analysis shown in this figure was performed by Karolina Worf at the Helmholtz Institute in Munich. **A.** Protospacer alignments to a ribosomal gene unit on chromosome 21 (blue from left to right: 5'ETS, ITSs, 3'ETS; orange: 18S rRNA, 5.8S rRNA, 28S rRNA) extracted from sequencing reads of three human MNase digest libraries (L, M, S libraries). **B.** Percentage of gRNAs harbouring a PAM sequence 3' of the protospacer target site. This analysis was performed for different published PAM sequences specific to particular Cas9 variants as shown. **C.** Percentages of gRNAs falling onto autosomes and sex chromosomes. Distribution of genes and DNA bases are shown for comparison. **D.** Theoretical number of gRNAs with *S. pyogenes* PAM sequences accessible to CRISPR EATING and MNase digest in the *E.coli* genome. **E.** GC content of the different gRNA libraries relative to the GC content of the genomes they were generated from. **F.** Analysis of nucleotide composition around the endonuclease target sites. Labels "First" and 'Last" denote the 5' and 3' ends of the gRNA that is cloned into the library vector -1 is the position upstream of the genomic cut site from which the gRNA is excised. +1 is immediately downstream of the cut site and is the first position of the PAM sequence.

3.3 Discussion

3.3.1 dCas9-chromatin modifier constructs

I successfully constructed a toolbox of targetable chromatin modifier proteins for transient expression. I selected a subset of constructs, for which similar constructs had in the meantime been published and shown to work by other labs, to establish monoclonal cell lines expressing these selected constructs. I chose dCas9-p300, dCas9-TET1 and because it is also associated with gene activation, dCas9-SET7 (although this construct has not been validated to date).

In hindsight, I should have more carefully characterised the monoclonal cell lines I established (see section 5.2 on Troubleshooting). At the time, I was convinced that based on the available literature the presence of the T2A-Blasticidin sequence on the plasmid would ensure that the dCas9-chromatin modifier is transcribed and translated, otherwise cells would not acquire Blasticidin resistance. Given that the cell lines were clearly resistant to Blasticidin, the most likely explanation why Cas9 expression was undetectable by Western blot in most of the monoclonal cell lines was due to the sensitivity of the assay. Another plausible explanation was that cells could selective degrade the dCas9-chromatin modifier protein once it had been made while keeping the Blasticidin resistance protein. Because I

reasoned that the problem was likely at the protein level (either having too little protein to detect or selective degradation). I did not perform genotyping or qPCR assays on the monoclonal cell lines at this point (see section 5.2).

3.3.2 gRNA libraries suitable for an epigenetic screen

I constructed several different gRNA libraries. These differ in their complexity, number of target sites, targeting efficiency, and coverage of the genome. A comparison to the gRNA library used in a published CRISPR-based transcriptional activator screen [92] is shown in **Figure 3.19**. The EMT5000 library was constructed from custom oligonucleotides synthesised on chips and consists of just 5,000 different gRNAs targeting 15 genes known to be involved in the regulation of epithelial-to-mesenchymal transition (EMT). I also wanted to develop a new method for the generation of ultra-complex libraries that is cost-effective. Custom oligonucleotide pool synthesis is currently limited to the synthesis of roughly 100,000 different sequences per chip, is expensive and requires special equipment (oligonucleotide synthesizer) not commonly available. Custom design libraries are suitable for screens where target regions are limited and can be clearly defined *a priori*. This applies to small scale epigenetic screens and also to genome-wide genetic screen, where an indel leading to a frameshift or premature stop anywhere within one of the 20,000 known protein-coding sequences will produce the desired knockout. Likewise, in the case of the activator/inhibitor screen guides were targeted just upstream (activator) or downstream (repressor) of the transcriptional start site [92]. However, for an epigenetic screen it is less clear which site needs to be targeted with the chromatin modifier to exert an influence on transcription. From the studies published to date that tested targetable chromatin modifiers at individual loci it is clear that often a subset of sites in relatively large promoter or regulatory regions exert a functional effect, with neighbouring sites producing no effect. For example, a gene promoter may have hundreds of CpG sites any one of which may be functional. It is feasible to comprehensively cover larger regulatory regions for a handful of genes in a small-scale epigenetic screen, and libraries for such approaches can be generated using standard *in silico* design combined with oligonucleotide pool synthesis. When the aim is to interrogate many more sites, costs of synthesis quickly become prohibitive. In order to be able to carry out genome-wide epigenetic screens in the future, ultra-complex gRNA libraries will be required and easy, cost-effective methods for generating these are highly desirable.

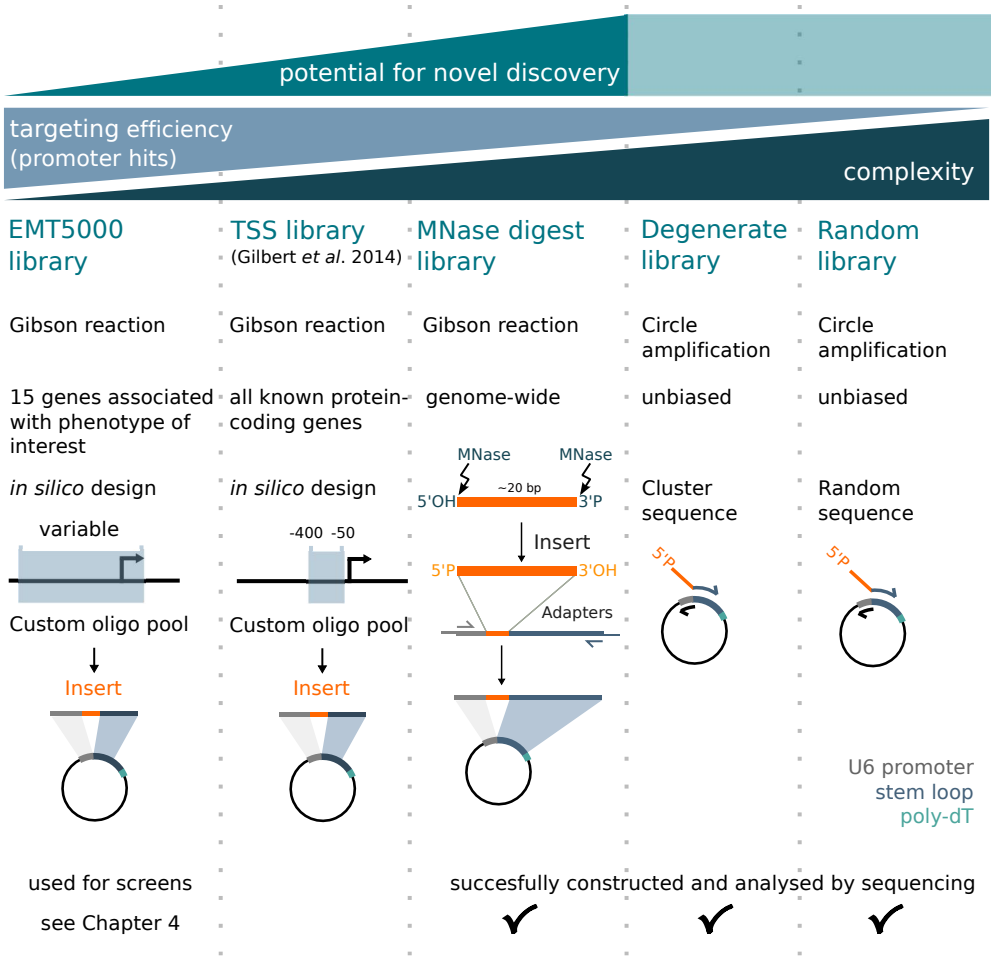


Figure 3.19: Comparison of the different gRNA libraries. Libraries differ in their targeting efficiency (defined as number of target sites divided by the total number of different sequences in the library, light blue bar), complexity (number of different sequences in the library, dark blue bar) and potential for discovering novel sites involved in the pathway of interest (turquoise bar). For the EMT5000 library, all unique 20 bp gRNA target sites tiling along the promoter regions (grey box) of 15 genes known to be involved in EMT (from ref. [123]) were found and synthesised by oligonucleotide pool synthesis. The target sequences were made into a library by cloning into the gRNA-pLKO.1 vector using Gibson cloning. A library focused to the -400 to -50 region from the Transcriptional start site of all protein-coding genes has been published by Gilbert *et al.* (TSS library)[92]. The MNase digest library was derived by digesting genomic DNA to roughly 20 bp fragments using micrococcal nuclease (MNase), followed by end repair, adapter ligation and cloning into gRNA-pLKO.1. A degenerate library was constructed by deriving degenerate, consensus oligonucleotide primer sequences that are biased towards binding in promoter regions. The library was made by using the degenerate oligonucleotide as a primer in circle amplification. The random library was made using a GN_{19}-targeting sequence.

3.3.3 How useful might ultra-complex gRNA libraries be in future?

As described in the previous sections the random and degenerate libraries have a very low targeting efficiency as they contain many gRNA sequences that do not map to the genome at all (at least when not allowing any mismatches in the alignment). In practice, however, gRNAs can bind target sites with mismatches to the protospacer sequence, which underlies off-target effects [140], which means that some gRNAs without genomic targets might be able to bind to sites. Ultimately, the usefulness of these libraries will thus have to be established in a screening experiment.

I decided to focus on the EMT5000 libraries for my initial screening experiments and only used the degenerate libraries as negative controls in the candidate validation experiments (see section 5.1). The MNase digest libraries appear to be more useful for use in screening experiments, given that they are generated from genomic DNA using enzymatic digestion. Thus, all gRNAs should have targets in the genome. However, not all of these are followed by a PAM sequence and only about 25 % harbour a canonical *S.pyogenes* PAM. The recently published CRISPR-Eating method aims to ensure the presence of a PAM 3' of the gRNA target site by using restriction enzymes with the recognition sites C/CGG, CC/NGG, and C/TAG [122]. However, this is only partly successful with only 44 % of gRNAs from their *E.coli library* actually targeting sites followed by a PAM. For both methods, the absence of PAM sequences downstream of a large number of target sites is a limitation. It is however, at present, hard to imagine how this could be avoided. As the CRISPR-Eating method has so far only been applied to the *E.coli* genome and selected PCR amplicons, the human and mouse MNase digest libraries represent the first large-scale gRNA libraries generated for a mammalian genome.

A potential limitation of the MNase digest libraries is the fact that the length of the cloned gRNAs ranges from 18-45 bp, with the median length of 28 bp. There is currently no evidence to suggest an upper limit to the length of a gRNA in the literature. Ran *et al.* showed that gRNAs 30 bp in length are functional [138]. I now provide evidence that gRNAs up to 46 bp can still direct wild-type Cas9 to its target site. Ran *et al.* showed that 30 bp gRNAs are trimmed down to 20 bp in cells. I therefore also included controls representing the short 20 bp versions of the long gRNAs. In all five cases, the short gRNAs were slightly more effective at inducing targeted indels. The observed trend might reflect the need to process

the longer gRNA, which may not always be successful [138]. However, given that I have shown that long gRNAs can still induce targeted mutations, this slight difference in efficiency should not compromise the usefulness of the entire library.

Given both the large complexity and the presence of PAM-less gRNAs in the MNase digest library, I would opt to increase the multiplicity of infection (MOI) used for infection of cells in a screening experiment. An MOI of 0.3 has been used in most screening experiments to date because this yields single copy integration of the gRNA expression cassette. Increasing the MOI leads to more cells with more than one integration event. This increases the number of false-positives but also reduces the number of cells that need to be infected in order to achieve necessary representation of the library. False-positives could further be reduced using a recursive screening strategy, that is performing multiple rounds of screening. After the first round, gRNAs would be extracted from cells of interest and used to prepare a new library of reduced complexity to use in a second round of screening.

Chapter 4

Results: Screening experiments

4.1 Establishing an EMT-related cadherin switch as a suitable readout

I sought to establish the screening method in a simple system and chose cadherin switching as a read-out. Cadherin switching, i.e. downregulation of E-Cadherin (encoded by *CDH1*) and upregulation of N-Cadherin (*CDH2*), occurs during epithelial-to-mesenchymal transition (EMT), a process thought to be involved in cancer metastasis. EMT can be induced in cells in culture by over-expressing transcription factors such as SNAIL, SLUG or TWIST [141–145]. Because of its reversible nature, it is thought that EMT cannot be brought about through mutation alone. Epigenetic mechanisms are thus thought to play a role in the regulation of EMT [146, 147].

Importantly, cadherin switching (whether through EMT or an EMT-like process) can be induced in cell culture through addition of a TGF-β-containing media supplement to cells (**Figure 4.1 A**). Three days after addition of the supplement, cells display marked changes in morphology (**Figure 4.1 B**) and switch from a rounded, cobblestone shape associated with an epithelial state (-EMT supplement) to a dispersed, spindle-shaped appearance with protrusion, associated with the mesenchymal state (+EMT supplement).

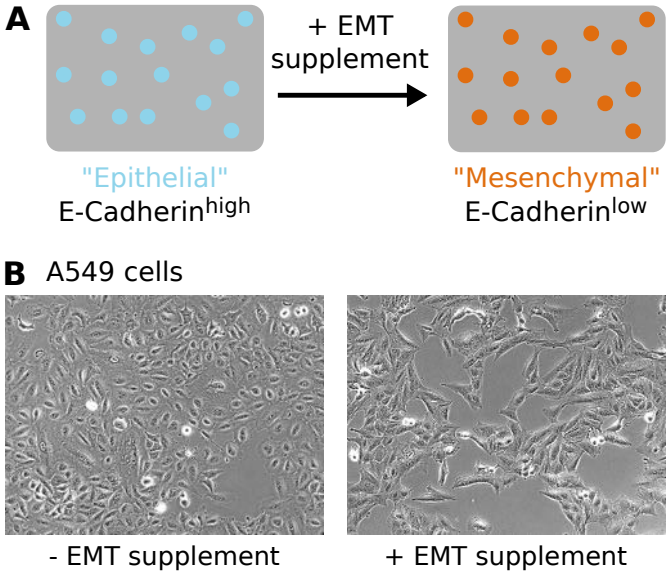


Figure 4.1: Establishing an EMT-related cadherin switch as a suitable readout. A. In the trial experiment, cells are cultured in the presence or absence of EMT-inducing media supplement. Addition of the media supplement is expected to induce the mesenchymal phenotype in all cells. **B.** A549 human lung carcinoma cells were cultured in the absence (left) or presence (right) of EMT-inducing media supplement for 3 days. Representative images from 3 experiments (DIC, 10X magnification).

A simple control experiment was conducted to establish the feasibility of using cadherin switching as a read-out: A549 cells were grown in the absence or presence of an EMT-inducing media supplement and expression of E- and N-cadherin is quantified by FACS. Initially, 26.5 % of cells express E-Cadherin, while only 5.8 % of cells express N-Cadherin. If grown in the presence of EMT-inducing supplement, 0.2 % express E-Cadherin and now 55 % express N-Cadherin (**Figure 4.2** top). This establishes that A549 cells can be induced to undergo EMT and that available antibodies are suitable for quantifying expression of E- and N-Cadherin. Formaldehyde fixation after antibody staining, necessary because at the time no FACS machine was available in a Biosafety cabinet to sort cells infected with lentivirus (see later chapters), does not impede detection of this cadherin switch (**Figure 4.2**, top versus bottom panel).

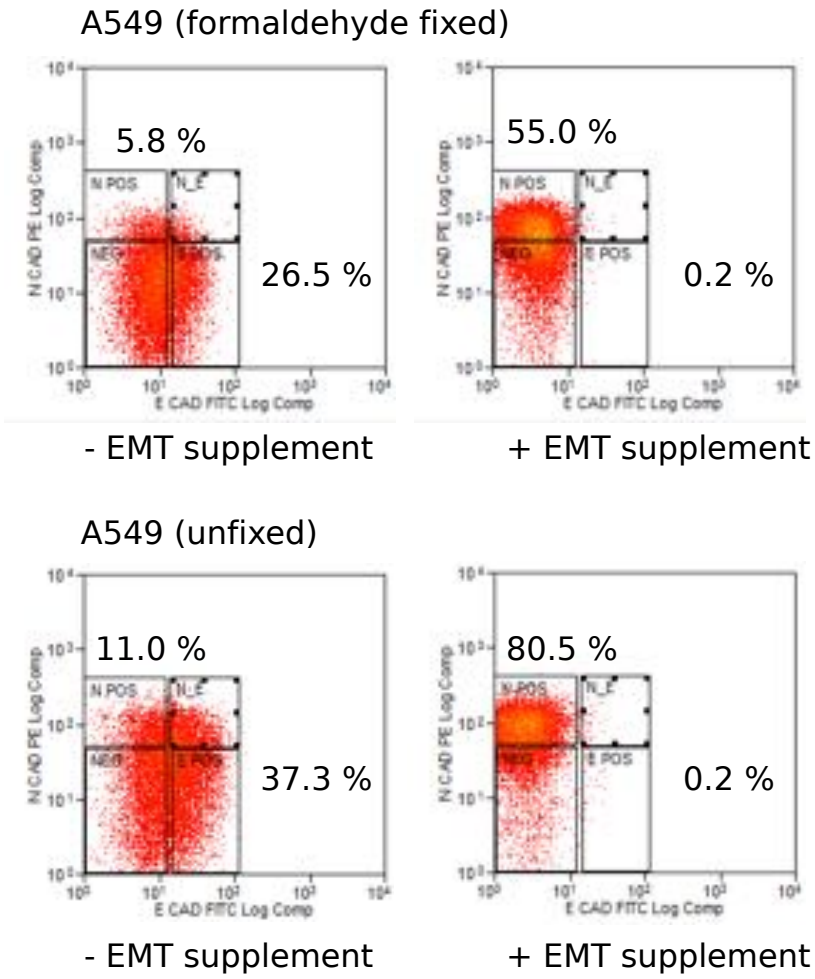


Figure 4.2: Cadherin switching as a read-out for an epigenetic screen. FACS analysis using antibodies against E-Cadherin (x-axis) and N-Cadherin (y-axis). A549 cells were grown in the presence (right) or absence (left) of EMT-inducing media supplement. To assess the effect of formaldehyde fixation on the assay, half the population was fixed with 10 % formaldehyde for 10 min before sorting (top versus bottom panel).

4.2 Screens with EMT5000 control library: Transient expression of dCas9 constructs

Next, a small-scale screening experiment was conducted using the EMT5000 control library. This library comprised just 5,086 gRNAs that tile across the promoters of 15 genes known to be involved in the regulation of EMT [123]. Few cells are expected to switch from E-cadherin to N-cadherin expression in the screening experiment as opposed to experiments using the EMT-inducing culture supplement. Not all gRNAs among the 5000 in the library will hit a functional site and bring about a change in gene expression strong enough to have a measurable

effect at the phenotypic level.

In this initial experiment, cells were transduced with the lentiviral library and then transiently transfected with one of several dCas9-chromatin modifier constructs as shown in **Figure 4.3**. Cells that displayed down-regulation of E-cadherin and concomitant up-regulation of N-cadherin (**Figure 4.3**, quadrant Q1) were identified by FACS.

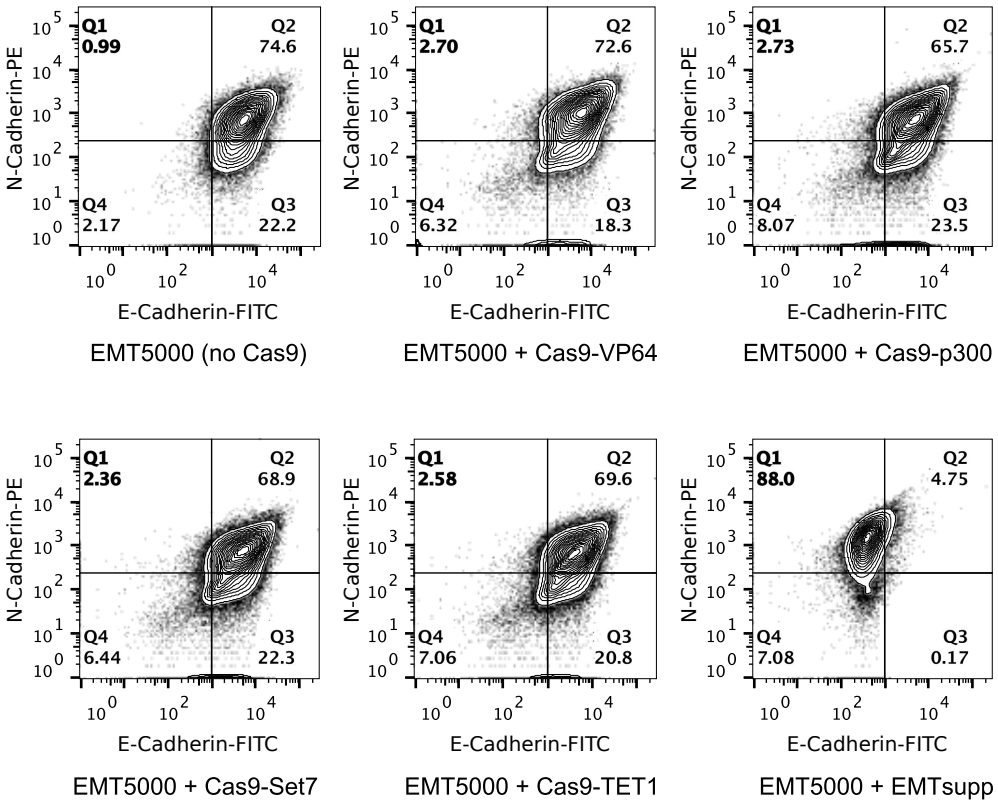


Figure 4.3: Screening experiment using the EMT5000 library and transient transfection of dCas9 constructs. FACS data from a screening experiment using the EMT5000 library and dCas9-chromatin modifier construct as indicated below the plots. Signal detected from E-cadherin antibody (FITC conjugate) is shown on the x-axis, N-cadherin (PE) is shown on the y-axis. The EMT5000 alone sample (top left) serves as a negative control, the EMT5000 alone sample grown in the presence of EMT-inducing media supplement (bottom right) serves as a positive control in the assay. Cells from Q1 were sorted.

Given that a similar shift was observed with all chromatin modifiers (although one might expect TET1 to show a weaker effect than p300, which directly recruits components of the transcriptional machinery), I was worried that the observed shift was solely a consequence of cell stress due to transient transfection. I was

further concerned that heterogeneous expression of the chromatin modifier may introduce a lot of variability and given the first genetic CRISPR-based screens [93–95] which were published around that time all used monoclonal stable cells lines expressing wild-type Cas9, I decided to also establish monoclonal stable cell lines expressing the targetable chromatin modifiers.

4.3 Screens with EMT5000 control library: Stable expression of dCas9 constructs

Monoclonal stable cell lines expressing a dCas9-chromatin modifier-T2A-Blasticidin construct were established as described in section 3.1.3. I chose to use the clonal lines expressing the dCas9-TET1 (Clone 13 and Clone 17), dCas9-p300 (Clones 12, 19, and 20) and dCas9-SET7 (Clones 1, 3, and 4) for these screening experiments because these chromatin modifiers are associated with gene activation and most gRNAs in the EMT5000 library target the promoters of genes that have been shown to induce EMT when over-expressed. The cells were transduced with lentiviral particles containing the EMT5000 library at an MOI of 0.3. This means 75 % of cells will not receive a construct and will be killed by Puromycin selection. Around 25 % of cells should have been transduced with at least one virion, the vast majority of transduced cells will have only a single gRNA integration event. I seeded $4x10^5$ cells the day before transduction. Assuming that cell numbers increased to at least $6x10^5$ cells at the time of transduction this puts the number of successfully infected cells at $1.5x10^5$, and given the library complexity of 5,086 gRNAs this yields approximately 30-fold coverage of the library.

For FACS, cells were detached with Versene and stained for expression of E- and N-cadherin. Versene was chosen because other cell detachment reagents such as trypsin would cleave cadherins on the cell surface. Clonal lines varied a lot in their basal expression level of N-cadherin and quite a lot of cells appeared to be expressing N-cadherin already in the negative control samples without library (see **Figure 4.4** for one of the most pronounced cases). Therefore, I decided to sort cells solely based on expression of E-cadherin (from gate P6 as shown in **Figure 4.4**).

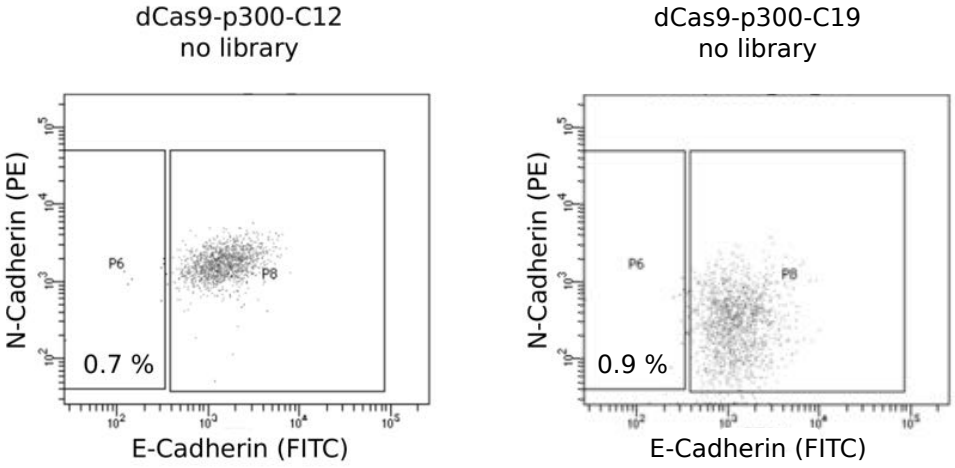


Figure 4.4: Staining of monoclonal dCas9-p300 cell lines for expression of E- and N-Cadherin. FACS data from A549 cells stably expressing the dCas9-chromatin modifier constructs indicated. Signal detected from E-cadherin antibody (FITC conjugate) is shown on the x-axis, N-cadherin (PE) is shown on the y-axis.

As predicted, far fewer cells undergo a cadherin switch in the screen compared to the cells grown in media containing EMT-inducing supplement (**Figure 4.5 A** and **B**). This is to be expected because few (if any) gRNAs from the EMT5000 library will elicit an effect strong enough for cells to undergo EMT. Gates were chosen so as to sort less than 1 % of cells in the negative control sample (see gate P6 in **Figure 4.4**) and cells that had lost E-cadherin expression in the populations that expressed both the gRNAs and dCas9-chromatin modifiers were isolated by FACS sorting. Cells were also sorted from the "library only" negative control (no dCas9-chromatin modifier). The percentage of E-cadherin negative cells that were sorted from the population in these screens are summarized in **Figure 4.5 C**. Cells were sorted into Quick Extract DNA extraction solution and following heating steps to promote cell lysis, used directly as input for PCR to amplify gRNA sequences for preparation of a next-generation sequencing library.

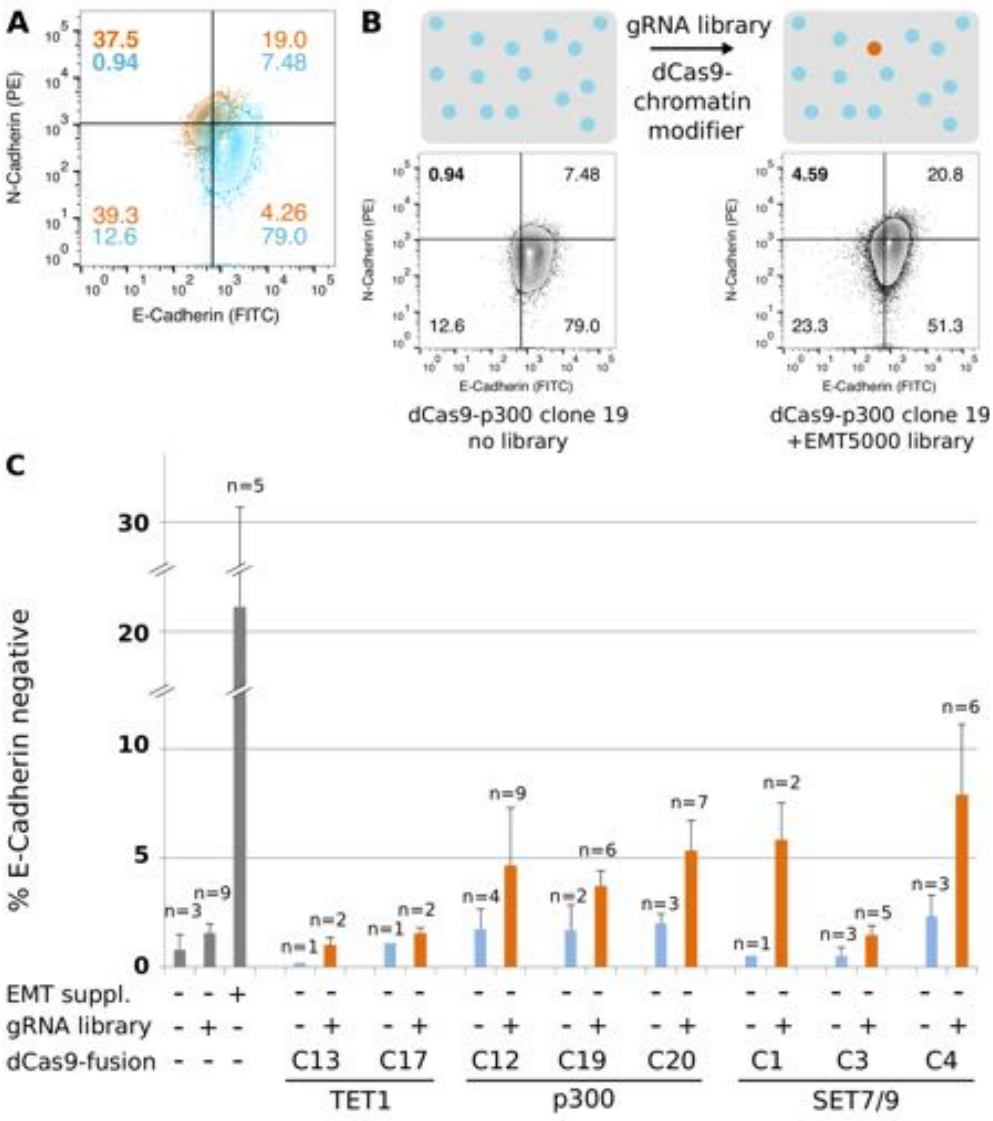


Figure 4.5: Screening experiments using the EMT5000 library and stable expression of Cas9 constructs. Signal detected from E-cadherin antibody (FITC conjugate) is shown on the x-axis, N-cadherin (PE) is shown on the y-axis. **A.** Untransduced controls cultured in the absence (blue) or presence (orange) of EMT-inducing cell culture supplement. **B.** In the screening experiment few cells in the population are expected to switch from an E-cadherin-high to an E-cadherin-low phenotype because few (if any) gRNAs in the library comprising 5,000 are expected to target a functional site where chromatin modification induces an effect strong enough to elicit an epithelial-to-mesenchymal transition. Representative examples of negative controls (no library) and the same clonal cell population after transduction with library are shown **C.** Bar graph showing percentage of E-cadherin negative cells that were sorted from the population of cells. n = number of technical replicates (same clonal line independently transduced with EMT5000 virus) across multiple different screening experiments. The different clonal lines (e.g. Clone 13 and Clone 17 for the dCas9-TET1 expressing cell lines) are considered biological replicates in these experiments.

4.3.1 Accurate counting of gRNAs from FACS-sorted cells

In order to identify candidate gRNAs it is necessary to compare the frequency of gRNA sequences extracted from FACS-sorted cells in the samples with those gRNAs sequenced in the negative controls. gRNAs that target the chromatin modifier to a functional site where modification induces the desired change in phenotype should be over-represented in the sample relative to the negative control. A particular challenge was to extract and sequence the gRNAs from approximately 10,000-30,000 cells sorted per sample. This number of cells is too large to handle with currently available single-cell analysis methods, which are limited to the analysis of a few hundred cells. Given that cells were infected at a low MOI, each cells likely contains a single insertion of the gRNA expression cassette. This means $4.98\text{x}10^{-20}$ mol target DNA sequence is present in each sample, which translates to approximately $5\text{x}10^{-15}$ g starting material. With current technology it is not possible to sequence femtograms of DNA directly without PCR amplification. I wanted to avoid PCR amplification because amplification bias would make it impossible to accurately count and compare gRNA numbers in samples and controls. Usually, PCR duplicates can be removed from sequencing data during analysis by collapsing reads that have exactly the same length and sequence (e.g. in a typical RNA-seq work-flow). However, this approach is not applicable here as the gRNAs are sequenced following targeted amplification of the gRNA sequence using primers that align to the U6 promoter and scaffold sequence respectively. This means that amplicons from two different starting molecules of identical sequence will be indistinguishable from PCR duplicates of a single starting sequence. Especially after many cycles of PCR amplification, raw read counts are an unreliable indicator of the number of starting molecules due to stochastic amplification bias [148].

In order to reduce the impact of PCR amplification bias I decided to attach unique molecular identifiers (UMIs) [149, 150] to the amplicon before amplification. These are random nucleotides introduced at a particular location in the sequencing read so as to uniquely barcode starting molecules. This strategy makes it possible to identify PCR duplicates in targeted amplicon sequencing. Instead of counting the number of reads associated with each gRNA, the gRNA count for each sample is derived by counting how many times a particular gRNA appears with a different UMI (**Figure 4.6 A**). When added by ligation or during the reverse transcriptase step (e.g. in single-cell RNA-seq library preparations), each UMI labels a unique original target molecule. Because the gRNA expression

cassette is integrated into the genome following viral integration, it is not possible to ligate an UMI in my case. Instead UMI barcodes are added by two (Option A) or three (Option B) initial cycles of PCR (**Figure 4.6 B**). In a second round of PCR both the gRNA sequences and UMIs are amplified using primers that anneal outside the UMIs. For the first round of PCR amplification, I designed primers that would anneal to the U6 promoter and gRNA scaffold sequences respectively and harbour overhangs consisting of an Illumina Nextera adapter sequence and a seven nucleotide (N7) UMI. I tested these primers and achieved successful amplification of a segment 175 bp in length that harbours the gRNA protospacer sequence in its centre (**Figure 4.6 C**). Amplification from genomic DNA extracted from A549 cells not transduced with the lentiviral library yielded no amplification products, showing that amplification is specific to the gRNA cassette.

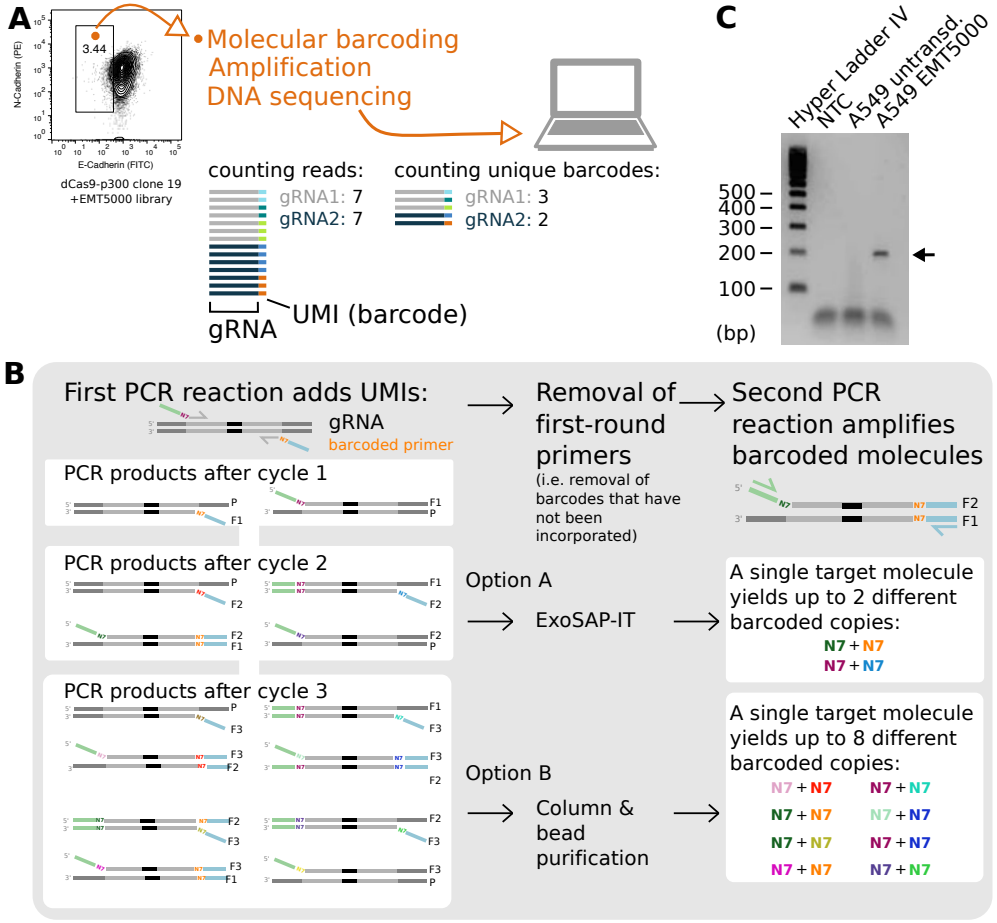


Figure 4.6: Barcoding PCR used for gRNA counting with NGS sequencing. A. Illustration of steps from FACS to counting of gRNAs in the samples. **B.** Details of amplification of gRNA sequences from DNA of sorted cells using primers with unique molecular barcodes (UMIs). **C.** Amplification of gRNA sequences using barcoded primers yields a band of the expected size (175 bp). No amplification products are seen after 34 PCR cycles in the no-template control (NTC) or using DNA from untransduced A549 cells as input.

As depicted in **Figure 4.6 B**, I tested two protocols for barcoding and amplification of gRNA molecules from sorted cells. In the first (Option A), two cycles of PCR with 1^{st} round N7-barcoded primers, which is the minimum number of cycles required to produce amplifiable product, are followed by digestion of unincorporated barcoded primers using ExoSAP-IT, a proprietary enzyme cocktail containing Exonuclease I and Shrimp Alkaline Phosphatase (SAP), which digests single-stranded DNA and dNTPs. These enzymes are then heat-inactivated and the barcoded molecules amplified using 2^{nd} round primers as shown. PCR mastermix and 2^{nd} round primers can be added directly to the reaction, bearing the advantage that this three-step amplification procedure can be performed in a

single tube without any intermediate purification steps. This minimises loss of sample, which is particularly important given the low amount of starting material.

It is difficult to ascertain that digestion with ExoSAP-IT is complete. If barcoded 1^{st} round primers were incompletely removed, new barcodes could be continuously incorporated during the second round of PCR and hamper accurate counting of gRNAs. While difficult to imagine how to detect the presence of a few remaining undigested primer molecules, I tried to devise a strategy to at least rule out that there is a lot of undigested primer left in the reaction following incubation with ExoSAP-IT. I performed two cycles of PCR with 1^{st} round barcoded primers, followed by addition of ExoSAP-IT or water and a second round of an excessive 48 cycles of amplification with non-specific primers. I had originally designed these primers for second round library preparation but found in initial testing (data not shown), that these primers, while designed against the Illumina Nextera adapter sequences, also amplify non-specific fragments from genomic DNA, including a 300 bp size fragment (**Figure 4.7 A**, lanes 4,6,8). It appears that in the presence of excess barcoded 1^{st} round primers, the 175 bp product is preferentially amplified, even in the presence of non-specific 2^{nd} round primers (**Figure 4.7 A**, lanes 2 and 3). Upon addition of ExoSAP-IT, only the 300 bp fragment is amplified (**Figure 4.7 A** , lane 4) but when the amount of ExoSAP-IT is reduced below recommended levels, both fragments are amplified, suggesting the presence of large amounts of incompletely removed 1^{st} round primers (**Figure 4.7 A** , lane 8). At the amount of ExoSAP-IT used in this protocol no amplification of the 175 bp fragment was detected (**Figure 4.7 A**, lane 4), and no fragments are amplified when adding Master mix but no primers to the reactions for the second round of amplification (**Figure 4.7 A**, lanes 5,7,9), suggesting that most of the unincorporated 1^{st} round primer is successfully removed using ExoSAP-IT.

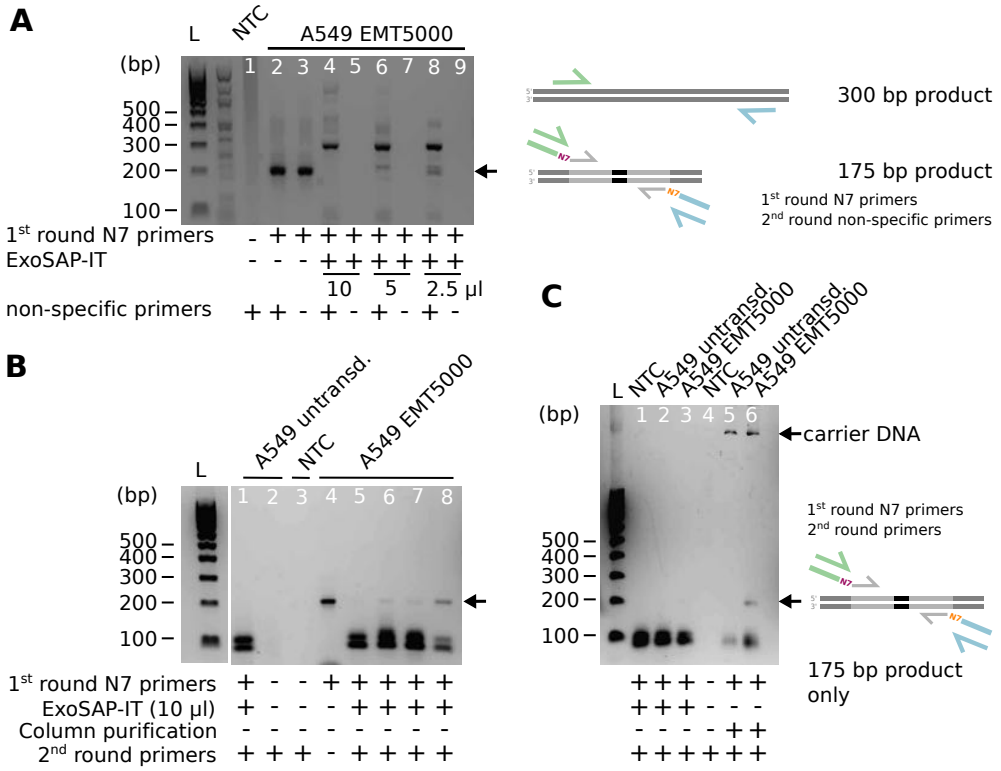


Figure 4.7: Sequencing library preparation. L denotes the Hyper Ladder IV (Bioline). **A.** Incubation with ExoSAP-IT digests most of the unincorporated 1^{st} round N7-barcoded primers. After two cycles of PCR with 1^{st} round N7-barcoded primers, ExoSAP-IT (or water) is added as indicated and reactions are incubated at 37 °C for 30 min. The enzyme mix is heat-inactivated at 80 °C for 20 min before addition of PCR master mix and non-specific 2^{nd} round primers (or PCR master mix and water) as indicated for a further 48 cycles of amplification. Non-specific 2^{nd} round primers may amplify both the gRNA cassette as well as a 300 bp fragment from genomic DNA as indicated in the cartoon on the right. **B.** Integrated gRNA protospacers are amplified from genomic DNA extracted from A549 cells transduced with lentiviral gRNA library or from untransduced control cells. A first round of amplification (two cycles) was carried out using N7-barcoded primers, followed by digestion of unincorporated primer using ExoSAP-IT, and a second round of amplification (34 cycles) using 2^{nd} round primers that further amplify the barcoded gRNA fragments. Annealing temperatures were varied in the second round of amplification in lanes 5-7 and were 69 °C, 65 °C, and 62 °C respectively. Annealing temperature for all other reactions were 65 °C. For the reaction in lane 8, 6 cycles of PCR were used for the first round of PCR and 30 cycles for the second round of amplification. The reaction in lane 4 was amplification with 1^{st} round primers only (34 cycles). **C.** Amplification from genomic DNA extracted from transduced and untransduced A549 cells as indicated. Three cycles of first round amplification with barcoded primers are followed by incubation with ExoSAP-IT or addition of carrier DNA and column and bead purification as indicated. This is followed by a second round of amplification (31 cycles).

PCR amplification for two cycles with barcoded primers, followed by treatment with ExoSAP-IT and a second round of amplification (34 cycles) yields faint but detectable amplification products 175 bp in size at different annealing temperatures (**Figure 4.7 B**, lanes 5,6,7). Increasing the number of cycles in the first round of amplification from two to six while keeping the total number of cycles constant increases the amount of final product, as expected. Amplification with 1^{st} round primers alone yields a product of the same size. Addition of these 2^{nd} round primers leads to the formation of two large primer dimers (compare **Figure 4.7 B** lanes 4 and 6). This appears to be a primer dimer rather than a non-specific amplicon because it is not produced when amplifying from genomic DNA of untransduced cells in the absence of 1^{st} round primers (compare **Figure 4.7 B** lanes 1 and 2).

I also tried a second protocol for amplification of gRNA sequences from FACS-sorted cells (**Figure 4.6 B**, Option B). Because this involved column and bead purification after the first round of PCR instead of digestion with ExoSAP-IT, I increased the number of cycles from two to three in the first round of amplification and included a carrier DNA (linearised eGFP-C1 plasmid DNA) in order to counter-act sample loss during purification. This approach also yields the expected 175 bp amplicon using genomic DNA from transduced but not from untransduced cells (**Figure 4.7 C** lanes 5 and 6). Following barcoding and amplification of gRNA sequence, indexed Illumina sequencing adapters were added by 6 cycles of PCR and fragments were sequenced on an Illumina HiSeq platform. A schematic illustration of the samples from the different screening experiments I prepared for next-generation sequencing are shown in **Figure 4.8**.

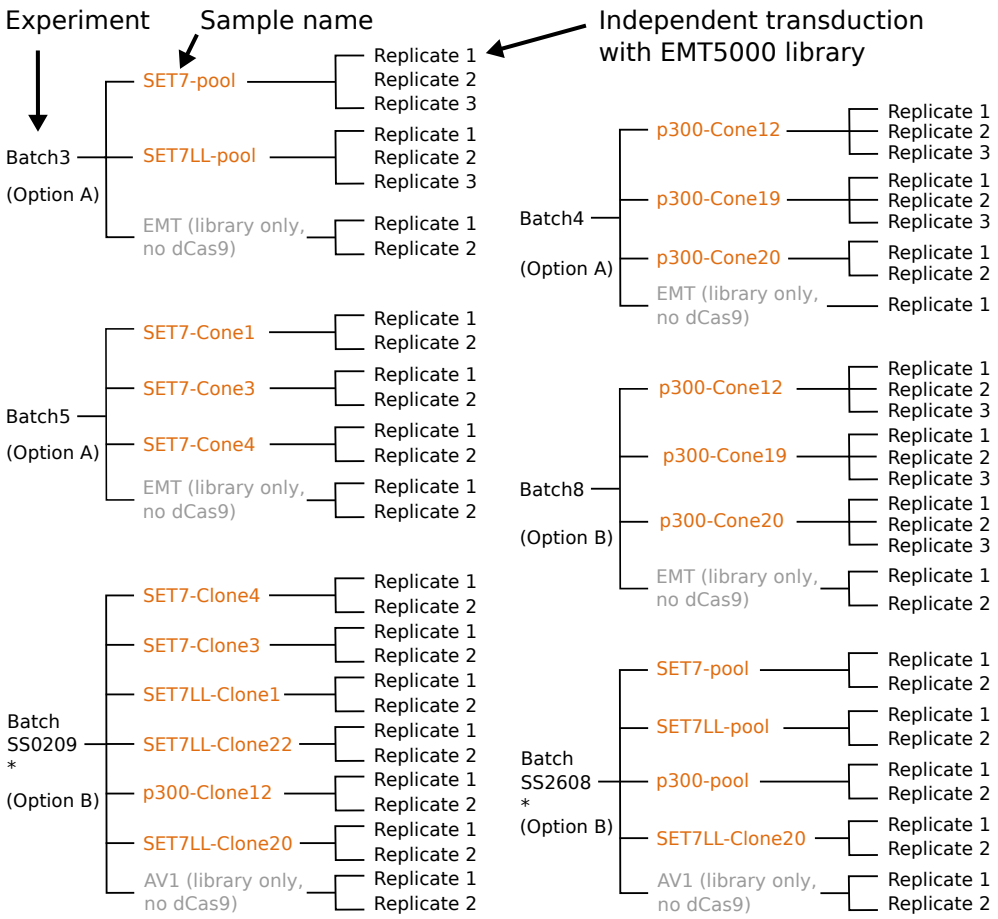


Figure 4.8: Overview of the samples prepared for next-generation sequencing. The different sequencing experiments are labelled as "Batches" with associated samples and controls shown in orange and grey respectively. The negative controls (grey) are A549 cells not expressing a dCas9-chromatin modifier that were transduced with the EMT5000 library sorted using the same gates as the samples. The negative controls are also referred to as the "library-only control". Whether two (Option A) or three (Option B) cycles were used for barcoding PCR is indicated below the Batch. Replicates denote technical replicates, i.e. independent culture, transduction with the EMT5000 library, staining, FACS sorting and sequencing. The two screening experiments labeled with a "*" (named BatchSS0209 and Batch SS2608) were conducted by my collaborator Stefan Stricker, who supplied pellets of sorted cells from which I prepared libraries for sequencing.

Guide RNA and UMI sequences were extracted from the sequencing reads and filtered for quality as described in the methods (section 2.5). Then a count was determined for each gRNA based on the UMIs rather than read counts. If two reads contain the same gRNA with the same UMI, these are treated as PCR duplicates. However, PCR does not just introduce amplification bias, it can also introduce errors. Without error correction, I found a linear relationship between

the number of reads per gRNA and the UMI-based counts of the gRNA (**Figure 4.9 B**, top). This is worrying because it suggests that the more a gRNA has been amplified during PCR, the higher its count despite using UMIs to correct for this bias. One possible explanation for this observation is that PCR error could be driving barcode diversity. The gRNA is identified by mapping the gRNA portion of the sequencing read back onto the EMT5000 library allowing mismatches (**Figure 4.9 A**). The UMIs from all the reads mapping to a particular gRNA are then used to derive the gRNA count. To test whether PCR (or sequencing) errors could increase barcode diversity and thereby artificially increase gRNA counts, I replaced the UMI with a different portion of the read and chose to use the uncorrected gRNA sequence. If there were no PCR or sequencing errors, I would expect to see a flat line. Even with thousands of reads for a particular gRNA the gRNA portion of the read should always be identical in the absence of errors, giving an overall gRNA count of 1 when these sequences are used as UMIs. However, this is not the case (**Figure 4.9 C**), suggesting that PCR or sequencing error introduces diversity into the gRNA portion of the read. The slope in **Figure 4.9 C** is not as steep as in **B**, which may reflect lower starting diversity in gRNA sequence (5,000 different gRNAs versus $2x4^7$ UMI sequences).

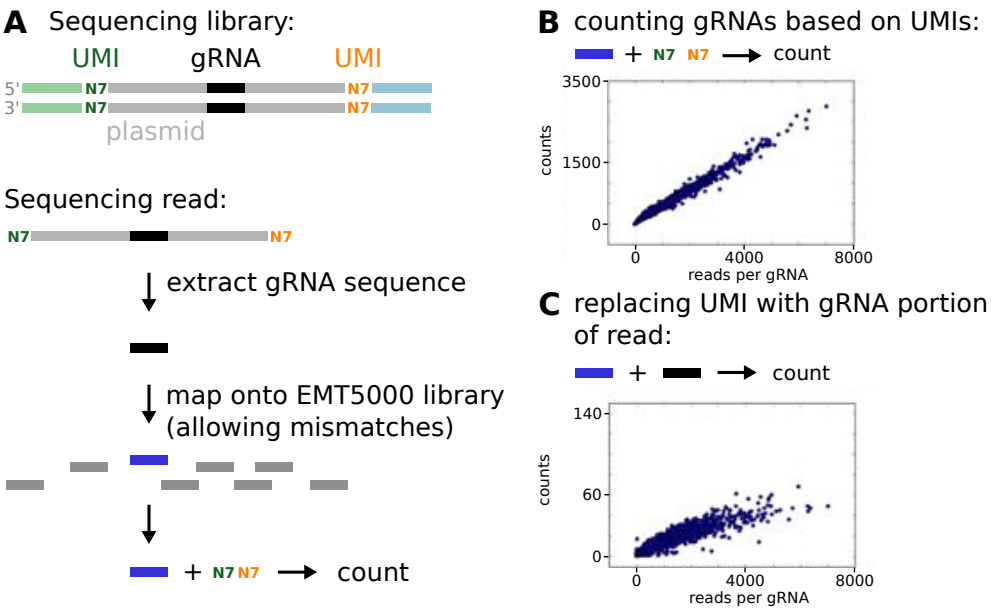


Figure 4.9: PCR error confounds gRNA counting. A. Schematic illustrating how gRNA counts are derived from next-generation sequencing data. gRNA sequences are extracted from the sequencing reads and mapped back to the reference EMT5000 library allowing mismatches to correct for PCR error. If a sequence does not map to the library the read is discarded. Next the 5' and 3' unique molecular identifiers (UMIs) are extracted from the sequencing read and combined. The gRNA count is the number of occurrences of different barcodes together with this gRNA. Reads harbouring the same gRNA and the same barcode sequences are identified as PCR duplicates and do not add to the gRNA count **B.** Without correcting for PCR error in the barcode, there is a linear relationship between number of reads and counts for the different gRNAs. Each gRNA is represented by a blue dot. **C.** Mapping to the EMT5000 library while allowing mismatches corrects for PCR or sequencing error in the gRNA portion of the read. If the uncorrected gRNA portion of the read is used for gRNA counting instead of the UMI, there is also a linear relationship between number of reads and counts for the different gRNAs.

The finding that PCR amplification introduces so many errors is surprising given that I used a high-fidelity polymerase for amplification. However, this might be due to the large number of cycles that had to be used for sequencing library preparation due to the low amount of starting material resulting from stringent selection of cells by FACS. It is thus important to correct for PCR error. A simple error correction, e.g. employed in most recent version of tools for the analysis of T-cell receptor repertoire sequencing [151], would be to group barcodes together up to a maximum number of allowed mismatches. The cutoff used in this method is somewhat arbitrary and does not take into account the fact that in my case, the UMI consists of the 5' and 3' molecular barcodes, which were ligated to the gRNA molecule by PCR independently of one another (see **Figure 4.6 B**). My

collaborator James Barrett developed a method for PCR error correction that is based on a Bayesian model (see Methods section 2.5). In brief, this model takes into account all the observed N7 barcode pairs for a gRNA and identifies the most likely number of latent barcodes, that is the underlying gRNA count that best describes the observed UMI sequencing data. Bayesian PCR correction was performed for all samples. A representative example is shown in **Figure 4.10**, which depicts scatter plots of gRNA counts versus reads for sequencing libraries generated from one screening experiment using the dCas9-p300 chromatin modifier. Without error correction, there is a strong correlation between the number of reads and count for each gRNA (left column). Grouping barcodes based on similarity with an arbitrary cutoff of 4 edit distances (mismatches) does not break this association (middle column). A Bayesian model successfully corrects for some of this bias (right column). After Bayesian error correction gRNAs with high number of reads also have low UMI-based counts. (Note that gRNAs with very low numbers of reads can never have high counts.)

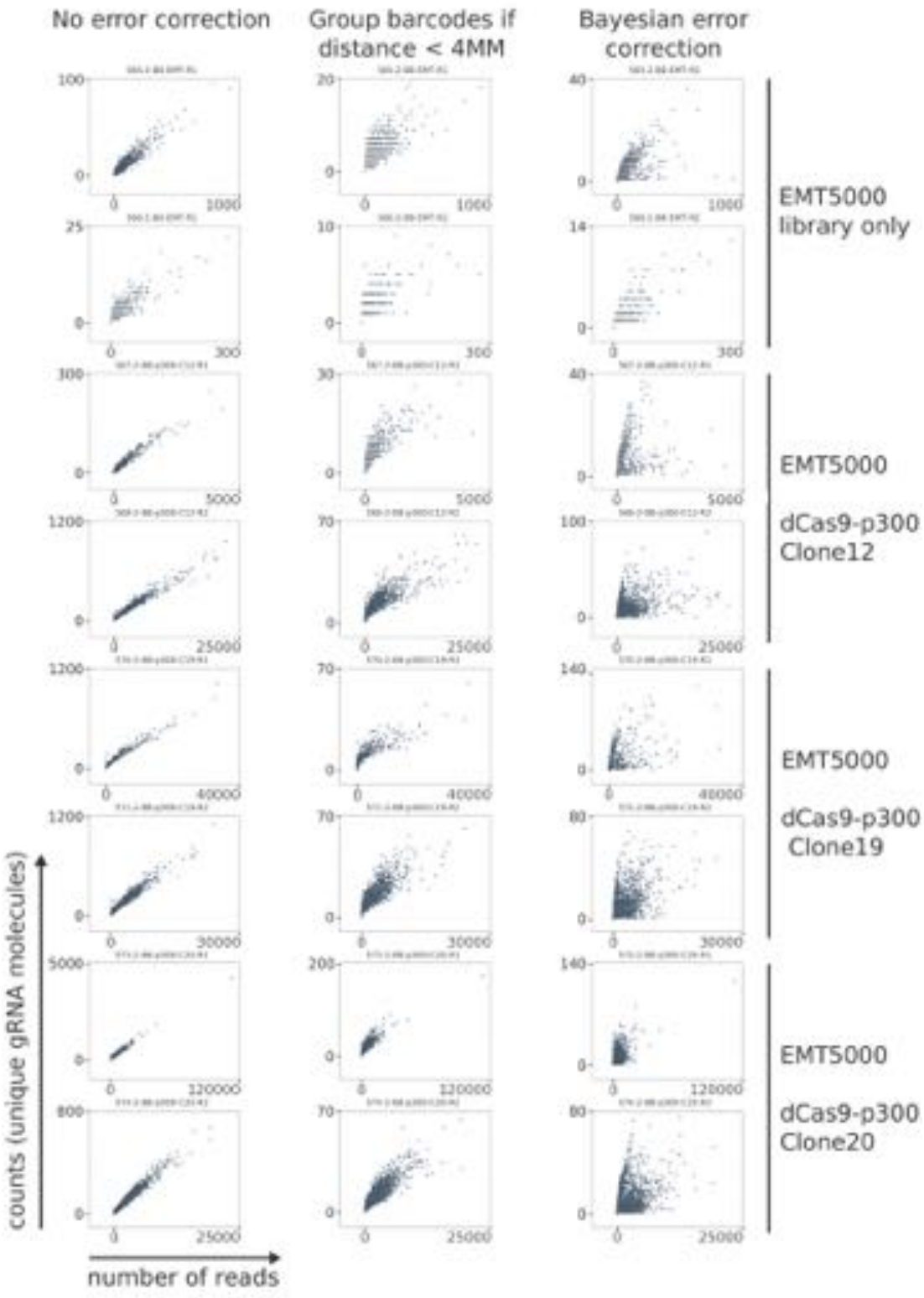


Figure 4.10: PCR error correction.

Figure 4.10: PCR error correction (continued). Scatter plots of counts for a gRNA versus the number of sequencing reads for the same gRNA without correcting for PCR error (left), when grouping barcodes together that are related by less than 4 edit distances (middle) or using a Bayesian model for error correction (right).

After successful PCR error correction, the total number of counts from each sample should reflect the number of FACS-sorted cells. A scatter plot of number of sorted cells versus number of total counts (calculated after Bayesian error correction) for each sequenced sample is shown in **Figure 4.11**. The observed counts may be lower than the recorded number of sorted cells for a given sample for several reasons: Not all cells are always successfully sorted or lysed, not all gRNAs are successfully barcoded in the first round of amplification and therefore lost (this will be worse for amplification Option A relative to Option B), molecules could be lost during purification (this will affect Option B more than Option A), or are not successfully amplified in the second round of amplification, or sequencing depth is insufficient.

However, the observed counts may also be slightly higher than the recorded number of sorted cells if barcoding of gRNA molecules is highly efficient (and this will affect Option B more than Option A). Using two cycles of initial barcoding, up to two barcoded molecules are generated for each starting molecule (**Figure 4.6 B**). With three initial cycles of barcoding, up to eight different barcoded molecules are generated, four of which are related and can potentially be collapsed by the error correction method. Furthermore, counts may be increased if barcodes acquire a lot of mutations that can no longer be corrected for or if additional barcodes are spuriously incorporated during later stages of the amplification reaction because barcoded primers were incompletely removed after the first round of amplification. Overall, I found that the total counts are lower than the number of sorted cells. Option A generally yields lower counts than Option B, which may reflect that fewer gRNAs will be successfully barcoded using just two PCR cycles of barcoding compared to three.

Ideally, there would be a method to directly determine the gRNA sequence integrated into the genome of the sorted cells without PCR amplification and sequencing library preparation. Both options I tried for barcoding gRNA molecules prior to amplification work, but from **Figure 4.11** it appears that Option B is more efficient than Option A, however, with the drawback of potentially over-counting gRNAs as discussed above (see **Figure 4.6 B**).

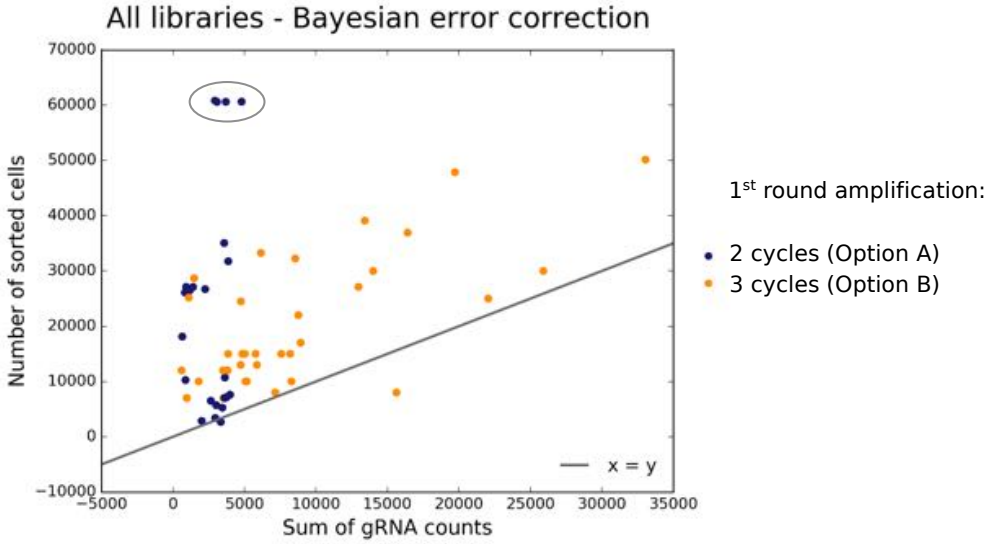


Figure 4.11: Scatter plot of number of sorted cells versus total gRNA counts per sequencing library. Only half the available lysate was used as input for generation of sequencing libraries in case of the four the outliers (grey ellipse).

4.4 Identification of candidate gRNA

I tried to test whether any gRNAs are significantly enriched in the samples relative to the library-only controls using DESeq2 [127], a software package originally developed to identify differentially expressed genes in RNA-seq data [152]. However, I found no significantly enriched gRNAs in my dataset. Using UMI-corrected counts means that the counts supplied to DESeq2 are much lower than when supplying read counts. DESeq2 should, however, be able to deal with this because sampling variance and overdispersion (extra within-group variability) are estimated from all the data. Then a test based on a negative binomial model is used to test the null hypothesis that there is no difference between sample and control groups. Thus as long as the differences between samples and controls are above the variance, DESeq2 should be able to detect changes. A computational pipeline for the analysis of genome-wide CRISPR-based loss of function screens called MAGeCK has also become available [153]. It uses the same method as DESeq2 to identify significantly enriched/depleted single gRNAs, however, then pools signal across multiple gRNAs that target the same gene to identify significantly enriched or depleted genes. This approach is however not applicable to epigenetic screens because neighbouring gRNAs cannot be assumed to have identical effects.

I was concerned that FACS, which constitutes a very strong selection, might introduce a lot of variance in my screening experiments. Gates were set so as to sort 1 % of cells in the library-only control (cells not expressing a chromatin modifier that were transduced with the EMT5000 library). When cells express both the targetable chromatin modifier and the gRNA, between 2-7 % of cells are sorted using the same gates. As expected the observed shift following epigenome engineering is small at the population level (see **Figure 4.5** above). Therefore, the rate of randomly selected false-positive cells is overall quite high. To date, only a single study reporting a CRISPR-based screen has used FACS to enrich for cells of interest [100]. However, no statistically significant hits could be identified in this saturating genetic screen at the *BCL11A* enhancer either. Similar to the analysis used by Canver *et al.*, I next tried to plot enrichment, i.e. Log2Fold change (Log2FC) calculated using DESeq2 for each experiment, along the chromosome, in the hope to visually identify regions in the promoters targeted by the EMT5000 library that would show increased signal compared to e.g. the *CDH1* promoter, which was included as a negative control. Example plots for four out of 15 genes targeted by gRNAs in the EMT5000 library are shown for one of the experiments (labelled "Batch8") using the dCas9-p300 construct are shown in **Figure 4.12**. However, unlike in the study by Canver *et al.*, this approach revealed no interesting candidate loci.

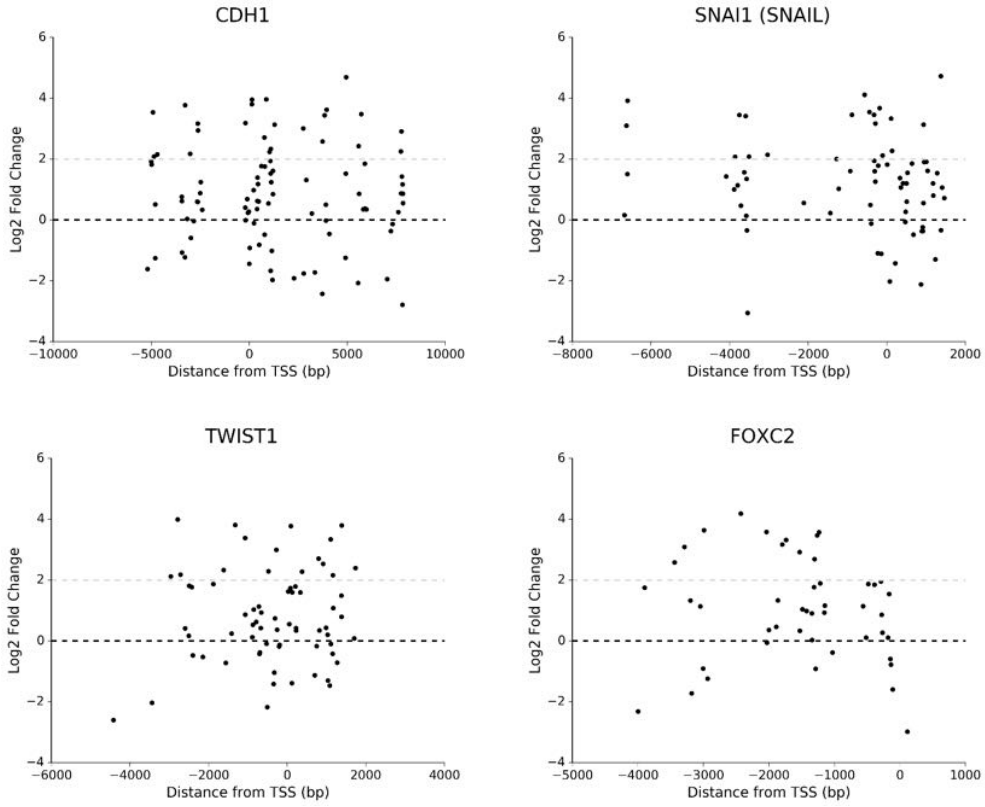


Figure 4.12: Log2FC along the chromosome for 4 out of 15 genes targeted by the EMT5000 gRNA library. Log2FC are from an experiment using the dCas9-p300 construct (experiment Batch8). Each dot represents the Log2FC calculated for a particular gRNA and is plotted onto its genomic location relative to the transcriptional start site (TSS) of the gene indicated.

Rather than trying to find those gRNAs that show the highest Log2FC compared to all other gRNAs, I next sought to identify those that show consistent enrichment across different experiments. As a first step, I performed a pairwise comparison of Log2FCs for the same gRNA across independent screening experiments. An example of a perfect correlation of Log2Fold changes across experiments is shown in **Figure 4.13 A**. I compared the three independent experiments that used the dCas9-p300 construct and EMT5000 libraries against each other in a pairwise comparison (**Figure 4.13 B**) and analysed the experiments using the dCas9-SET7 fusion construct in the same way (**Figure 4.13 C**). It is evident from these plots that the data is extremely noisy, with a similar number of gRNAs showing a negative correlation as show the desired positive correlation correlation in Log2FC scores between repeats of the experiments.

Given this low signal-to-noise ratio, I decided to try and identify candidate gRNAs using a simple ranking approach. I ranked gRNAs using an R package [128]

that implements desirability functions for ranking and prioritizing candidates in a variety of settings (**Figure 4.14**). The data values are first mapped onto a continuous scale from 0 to 1 (for a "high is good function") or from 1 to 0 ("low is good function"). I decided to prioritise gRNAs with a large positive Log2FC and low within-experiment variability (lfcSE, standard error on the Log2FC calculated using DESeq2). A weighted average is then calculated using data from all experiments to give an overall "Desirability" that can take values between 0 and 1. In generating this weighted average, Log2FC was given 4 times as much weight as standard error. This yielded a ranked list of gRNAs from which I chose the top ten for validation. As expected, the candidate gRNAs show higher UMI-corrected counts in the samples relative to the negative library-only controls (**Figure 4.15 and 4.16**).

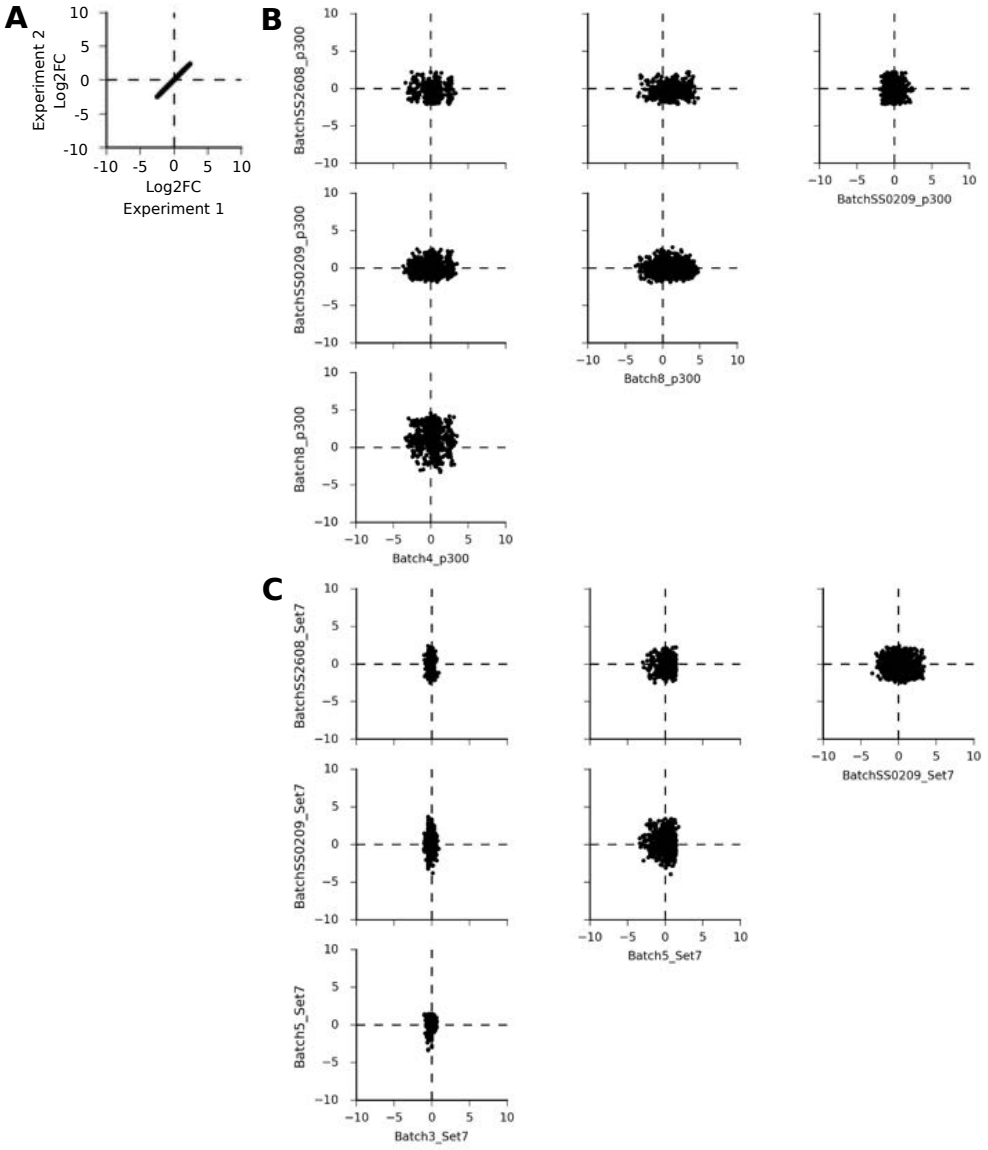


Figure 4.13: Correlation of enrichment (Log2FC) of each gRNA across different screening experiments. Each dot represents the Log2FC calculated for a particular gRNA. **A.** A perfect correlation is depicted for comparison. **B.** Scatter plots of Log2FC calculated for each gRNA using DESeq2 across four screening experiments using the dCas9-p300 construct (labelled "Batch4", "Batch8", "BatchSS0209" and "BatchSS2608"). **C.** Scatter plots of Log2FC calculated for each gRNA using DESeq2 across four screening experiments using the dCas9-SET7 construct (labelled "Batch3", "Batch5", "BatchSS0209" and "BatchSS2608")

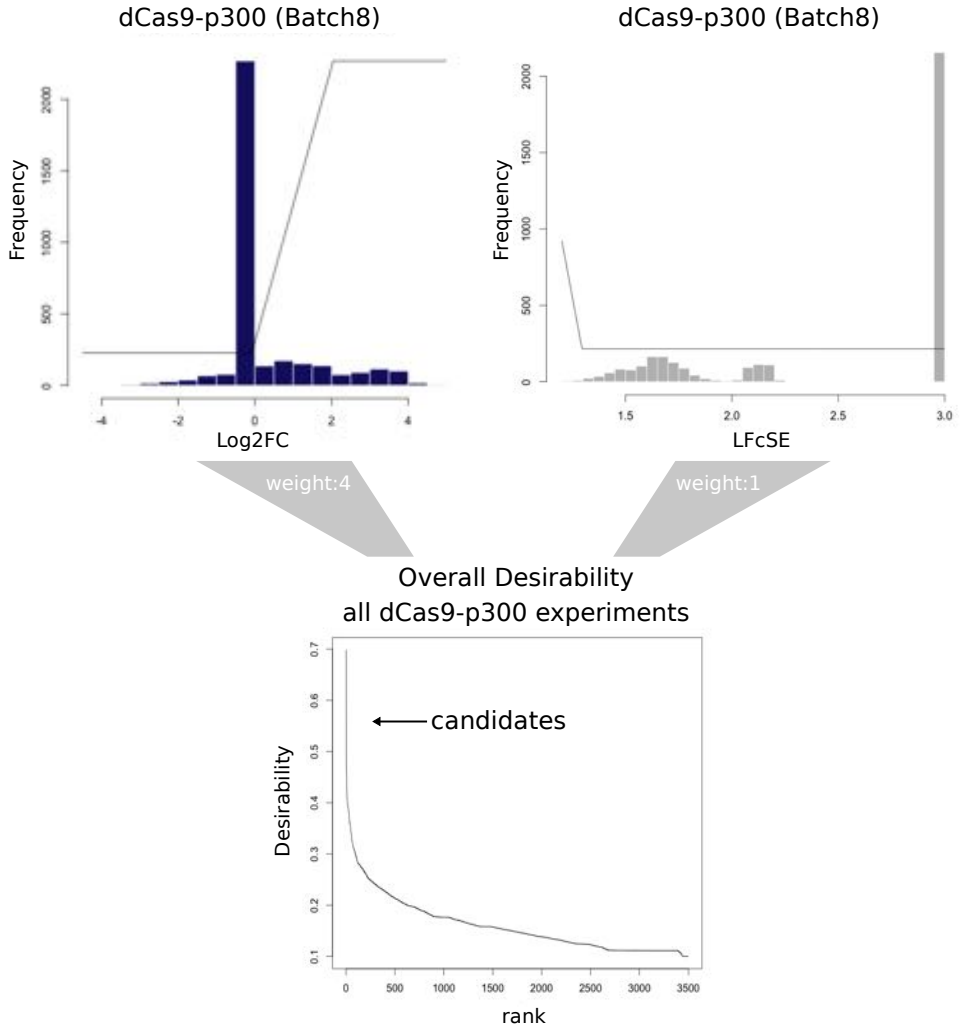


Figure 4.14: Selection of candidate gRNAs using a ranking approach illustrated by one example. Histograms of Log2FC (top left) and Log2FC standard errors (LFcSE, top right) calculated using DESeq2 are shown from one experiment using the dCas9-p300 construct and the EMT5000 library. These values are mapped onto a scale from 0-1 (indicated by the black line). These individual desirabilities are combined using a weighted average to calculate an overall desirability, which is used to rank the list of gRNAs (bottom). The top 10 highest ranking gRNAs were chosen as candidates for validation.

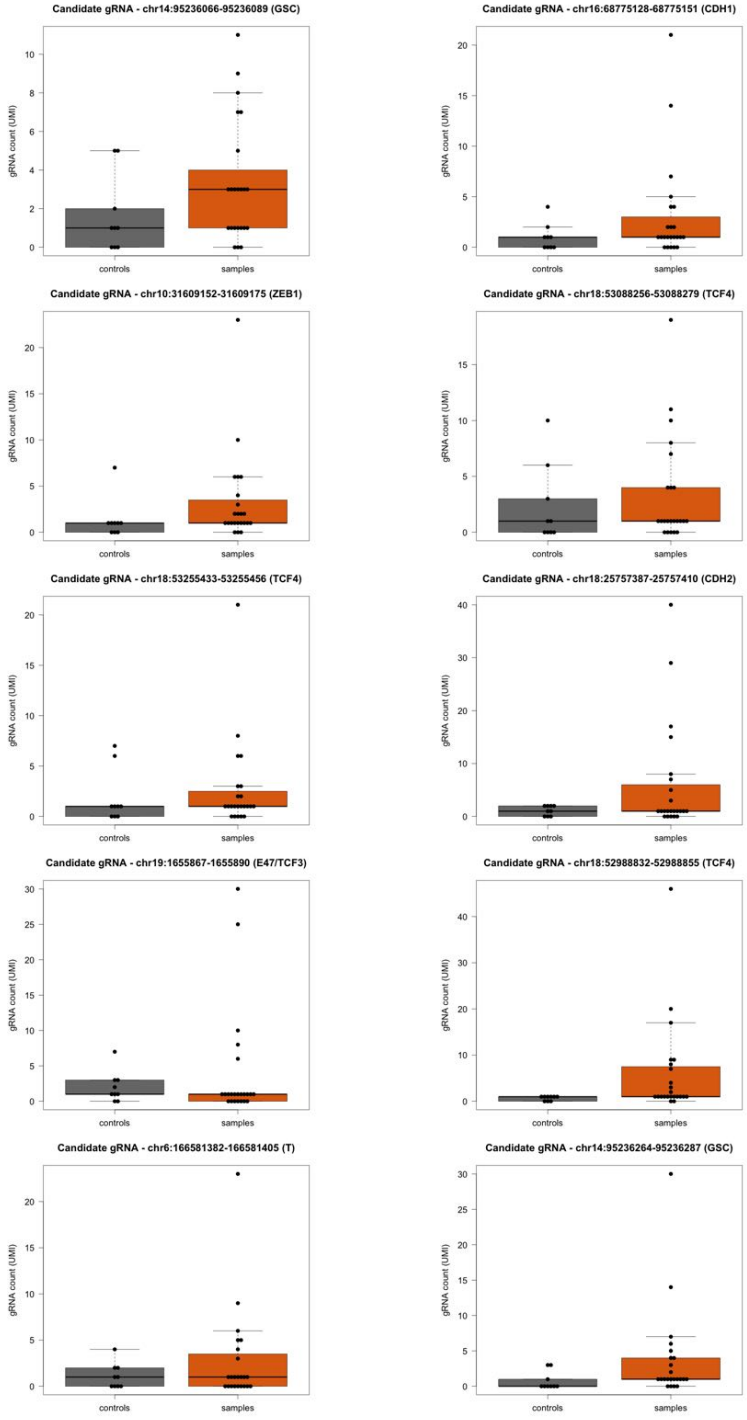


Figure 4.15: Top 10 candidate gRNAs from screening experiments using the EMT5000 library and the dCas9-p300 construct identified using a ranking approach. UMI-corrected counts are shown for each candidate gRNA for the no-library controls (grey) and samples across all screening experiments using this construct.

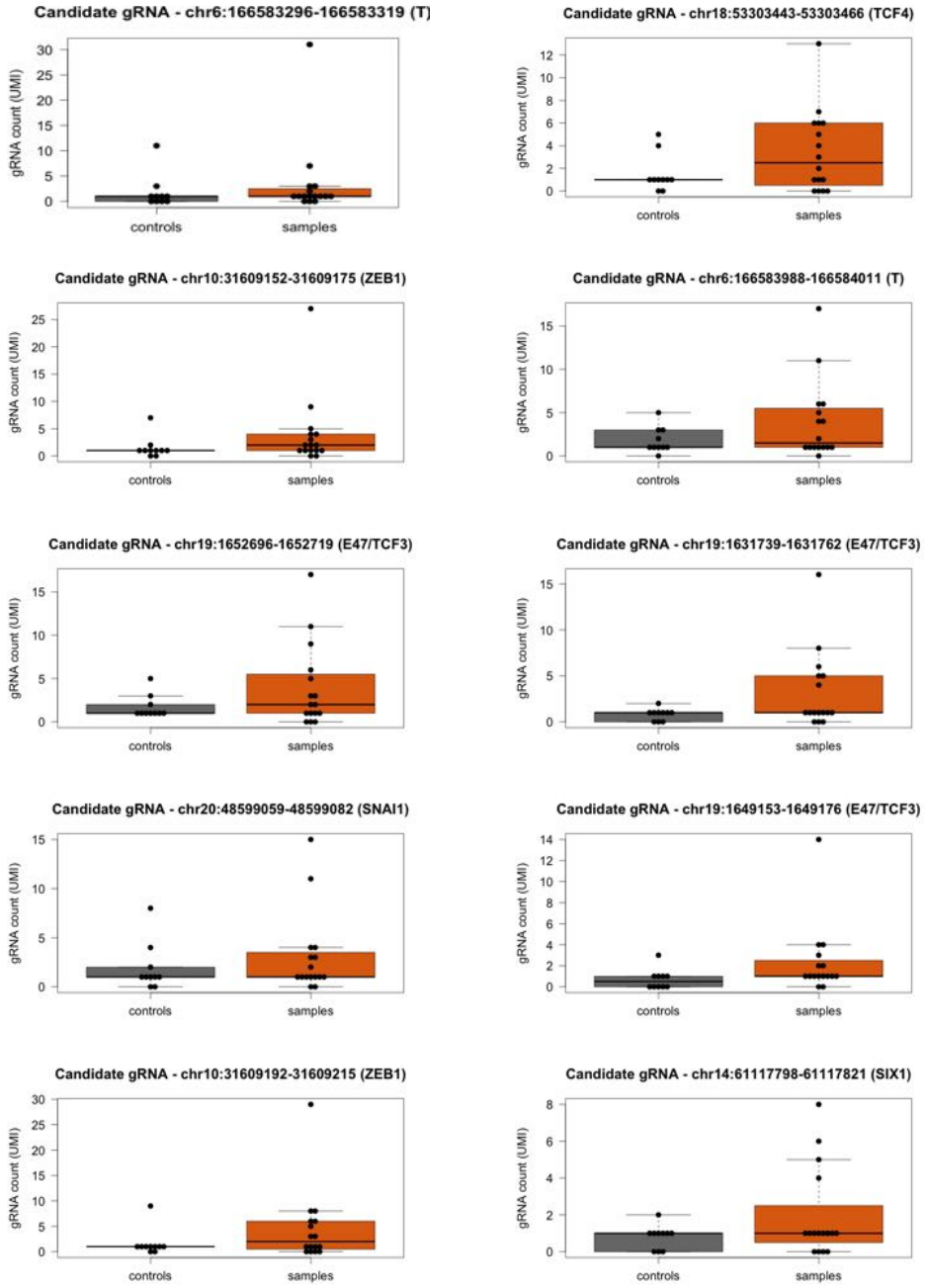


Figure 4.16: Top 10 candidate gRNAs from screening experiments using the EMT5000 library and the dCas9-SET7 construct identified using a ranking approach. UMI-corrected counts are shown for each candidate gRNA for the no-library controls (grey) and samples across all screening experiments using this construct.

4.5 Discussion

In this chapter, I introduced an EMT-related loss of E-cadherin expression on the cell surface as a possible phenotypic read-out for an epigenetic screen. A549 cells can be induced to undergo EMT *in vitro* using a cell culture supplement, which causes detectable changes in the expression of cell surface markers including E-cadherin. Despite problems with validation of dCas9-chromatin modifier fusion constructs (section 3.1.2), I decided to test the dCas9-p300, dCas9-SET7 and dCas9-TET1 constructs in a screening experiment using loss of E-Cadherin as a readout. These constructs were chosen because they are associated with gene activation and the majority of genes targeted by the EMT5000 library have been shown to induce EMT when over-expressed in cells. Furthermore, similar targetable TET1 and p300 constructs had been reported by other labs, suggesting that the catalytic domains of these proteins can be used for epigenome editing.

Using the cell culture supplement, essentially all (or most) cells in the population are expected to switch from an epithelial to a mesenchymal phenotype and lose expression of E-cadherin. In contrast, few cells in the population are expected to undergo the cadherin switch when using a targetable chromatin modifier together with the EMT5000 gRNA library in a screening experiment. Assuming gRNAs are evenly distributed in the synthesised library, 50 out of 5000 gRNAs would theoretically have to elicit a strong response in order to see a 1 % shift at the population level. Of course, gRNAs are known to be unevenly represented in the library following on-chip oligonucleotide synthesis, library and virus preparation. It is thus difficult to predict how much of a shift at the population level to expect in these experiments, but it is almost certainly smaller than the shift achieved using the cell culture supplement.

I detected small but reproducible loss of E-cadherin at the population level, presumably brought about by induction of EMT in those cells, using the targetable p300 and SET7 constructs, but not using targetable TET1 (**Figure 4.5 C**). Although a direct comparison between these targetable chromatin modifiers has not been reported to date, I would have generally assumed the TET1 DNA demethylase to be a weaker transcriptional activator compared to dCas9-p300, which has been shown to act as a strong activator [74], and is known to be able to directly recruit components of the transcriptional initiation machinery. Given this assumption and based on the observed loss of E-cadherin at the population level, I concluded that both the dCas9-p300 and dCas9-SET7 constructs were likely functional. I interpreted the lack of a shift in E-cadherin expression observed

with the dCas9-TET1 construct as the construct either being non-functional, or not able to elicit a measurable effect at the population level together with the EMT5000 library in this system.

For each repeat of the screening experiment, I tried to minimise the effect of variability in culture conditions and antibody staining. Control A549 cells and clonal lines expressing the dCas9-chromatin modifiers were transduced with EMT5000 virus in parallel, and maintained for several days at sub-confluent levels. Cells were then detached using Versene, counted and stained for E- and N-Cadherin expression in parallel, keeping the concentration and amount of antibody as well as incubation times and temperatures consistent. Dead cells were excluded from the analysis based on staining with a fixable viability dye (see Methods section 2.4).

In hindsight, additional controls should have been included in the experiment in **Figure 4.5 C**. It is possible that the shift in E-Cadherin expression observed at the population level is a consequence of infection of single-cell derived clones with lentivirus. Lentiviral infection may have a larger effect size in the single-cell derived clones compared to the "library-only" control, a polyclonal population of A549 cells not expressing any of the dCas9 constructs. This control was infected with the same virus in parallel. It would have been good to either use single-cell clones for the "library-only" controls as well or to include an additional "empty gRNA control", based on the library backbone vector without a gRNA protospacer sequence.

Convinced that the observed loss of E-cadherin in the screening experiments using dCas9-p300 and dCas9-SET7 was real, especially given that dCas9-TET did not induce such a shift, I decided to sequence the gRNA population in the samples expressing chromatin modifier and library and negative controls (expressing the EMT5000 library only, but no dCas9) from these experiments. One particular challenge was the low number of sorted cells (between around 5,000 - 30,000 cells per sample). One option to alleviate this problem would have been to scale up the experiment and grow more cells for FACS. However, it would have been necessary to increase the number of cells by at least 20-fold, which appeared not feasible, not least because of the cost of antibodies and time needed for sorting. Another option would have been to set less stringent gates for cell sorting, however, this would have also increased the number of false-positives in the sorted population. The rate of false-positives is already quite high (1 % of the population). I therefore opted for a strategy that is commonly used in single-cell

RNA-seq work-flows [154] and attempted to attach unique molecular identifiers (UMIs), sometimes also referred to as molecular barcodes, to gRNA molecules before PCR amplification. In RNA-seq protocols these UMIs are usually attached during the reverse-transcription step, or by ligation - both methods are inefficient but uniquely label individual molecules. This approach was however not applicable in my case. The gRNA expression cassette is integrated randomly integrated into the genome of transduced A549 cells and the only option to attach UMI appeared to be to use PCR with UMI-containing primers. This means that from each integrated gRNA, up to 2 barcoded molecules will be generated when using the minimum of two cycles of barcoding PCR. Using three cycles of initial barcoding, up to eight different UMI-gRNA combinations are generated for a single starting molecule. I tried both methods, two initial cycles may not efficiently label most gRNAs but introduces less variability, while three cycles produce more barcoded product. Sequencing libraries were successfully generated from lysates of sorted cells using both protocols.

The analysis of the sequencing data proved challenging. Perhaps epigenetic screens can be expected to have a lower signal to noise ratio in comparison to genetic screens, assuming that the effect of epigenetically modulating gene expression is probably not as strong as a genetic knockout. Most genetic screens performed to date have used sensitivity to a drug as a phenotypic readout. FACS has only been used to isolate cells of interest in a single study published to date [100], which also suffered from a lot of random noise. It is clear that cell sorting can introduce variability and false-positives. In addition, I had to use a large number of cycles of PCR amplification to generate enough material for sequencing library preparation and quality control given the low amount of starting material starting from low numbers of cells isolated by FACS. Both factors contribute to the observed noisiness of the data. Following filtering and PCR error correction of sequencing reads, I found no significantly enriched gRNAs in my datasets. In the end, I identified candidate gRNAs for validation using a simple ranking approach.

Chapter 5

Results: Candidate validation

5.1 Technical validation of candidate gRNAs

As described in the previous chapter, candidate gRNAs for validation were selected using a ranking approach. The protospacer sequences from the top 10 ranking gRNAs were cloned into the lentiviral gRNA vector gRNA-PLKO.1 as described in the Methods section 2.6. Cells expressing the targetable chromatin modifier were infected with the candidate gRNA and expression of E- and N-Cadherin was assessed by flow cytometry using the same staining protocol as for sorting of cells in the screening experiment, except for omission of the formaldehyde fixation step. (This step could be omitted due to better aerosol containment in the flow cytometer as opposed to the cell sorter.) Four negative controls were included in the validation experiment: (1) a gRNA against *GAPDH*, (2) a scrambled gRNA , and the degenerate libraries (3) C304 and (4) C422 (see **Table 3.1**). The rationale for also including entire libraries as negative controls was that a single gRNA could potentially (although this is improbable) bind an off-target site that induces the phenotype of interest. Cells were cultured in the presence of EMT-inducing supplement to provide a positive control that these cells can undergo a cadherin switch.

None of the candidate gRNAs induced an effect comparable to the culture supplement in the cells (**Figure 5.1**). In fact, E-cadherin levels stayed similar to those of cells transduced with the negative control gRNAs, indicating that none of the candidate gRNAs are functional. Three independent validation experiments using the SET7 candidate gRNAs and cells expressing the dCas9-SET7 chromatin modifier are shown in **Figure 5.1**.

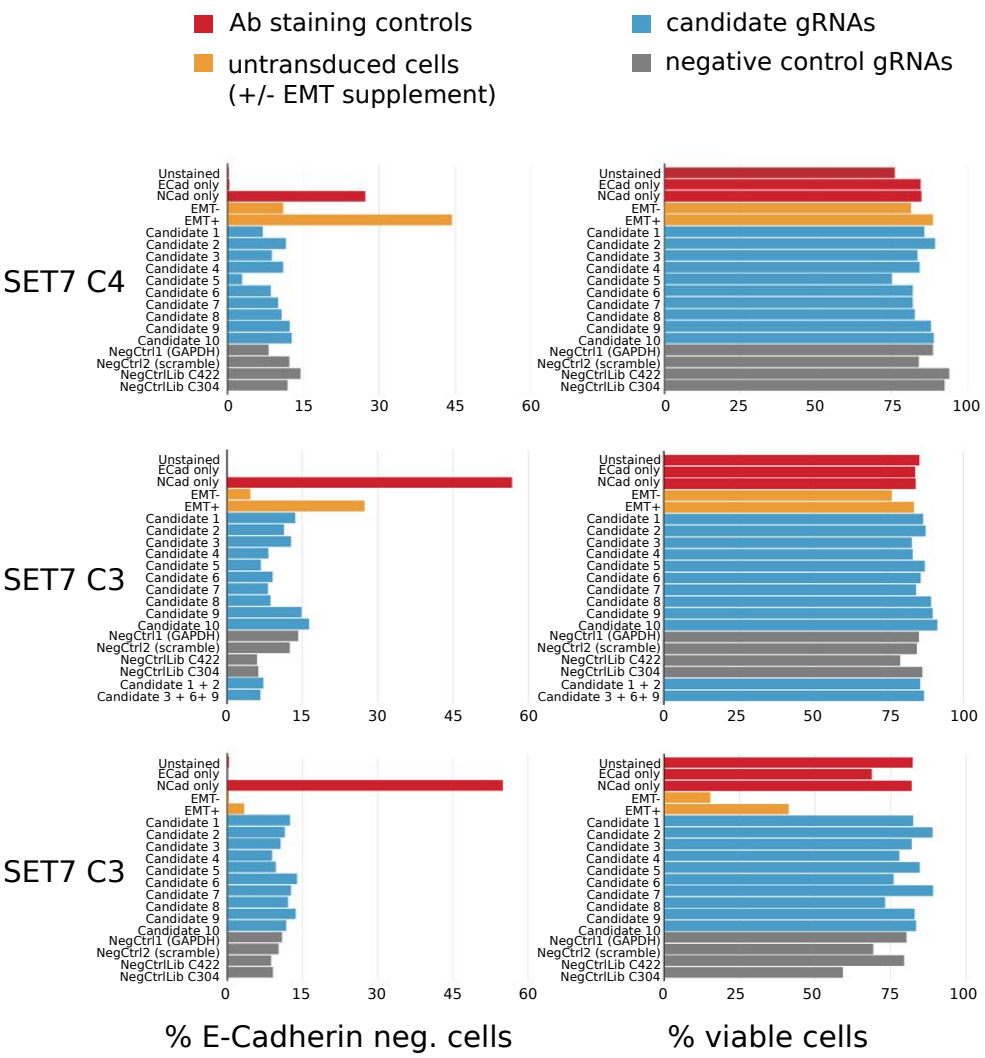


Figure 5.1: Technical validation of candidate gRNAs. A549 cells expressing the dCas9-SET7 chromatin modifier were transduced with candidate gRNAs or negative controls as indicated and E-cadherin expression levels measured by flow cytometry. The percentage of E-cadherin negative cells is plotted on the left and the percentage of viable cells is plotted on the right for each sample from the three independent experiments. Ab staining controls (red) include unstained and single-stained samples. Untransduced A549 cells cultured in the absence (-EMT) or presence of EMT-inducing culture supplement (+EMT) are shown in yellow. In one experiment (bottom panel), the untransduced control cells had low viability (due to a problem with detachment of cells in these two samples). Cells transduced with candidate gRNAs are in blue. Cells were also transduced with negative control gRNAs (grey); these included a gRNA targeting *GAPDH*, a scrambled gRNA, and two gRNA libraries made using degenerate oligonucleotides (Cluster 422 and Cluster 304).

5.2 Troubleshooting

Given that technical validation of the SET7 candidate gRNAs was not successful, I wanted to identify possible problems with the screening method and setup. As will be discussed in more detail below, there are several possible reasons why none of the candidate gRNAs identified by the screening method could be validated. Changes to the screening method could be made at almost every step of the protocol in an attempt to improve the method. In order to identify the most likely reason why the method did not identify reliable hits, it will be necessary to check every component of the screen including expression of the targetable chromatin modifiers and of the gRNAs from the library as well as to re-assess the suitability of the chosen phenotypic readout.

5.2.1 Integration and mRNA expression levels of dCas9-chromatin modifier constructs in monoclonal stable cell lines

I decided to first test the successful integration and mRNA expression of dCas9-chromatin modifiers from the monoclonal stable cell lines used in the screening and validation experiments. Only SET7-Clone3 was previously found to express the targetable chromatin modifier at levels detectable by Western blotting (**Figure 3.7**). However, cell lines for which protein was not detectable were clearly resistant to Blasticidin. Given that the resistance gene is located immediately downstream of the dCas9-chromatin modifier in one expression cassette and the encoded polypeptide is co-translationally cleaved at an intervening T2A peptide sequence, I had always assumed that the resistance gene cannot be expressed unless the dCas9-chromatin modifier is also successfully translated into protein. Furthermore, given the observed shift in the screening experiments I assumed the targetable chromatin modifier must be expressed. I now decided to test this assumption and given that Western blotting did not detect protein except in SET7-Clone3, I decided to check expression at the level of mRNA and also to amplify different parts of the expression cassette from transduced cells.

I designed primers tiling along the dCas9-SET7-T2A-Blasticidin and dCas9-p300-T2A-Blasticidin expression cassettes as shown (**Figure 5.2** and **Table 2.12**). Genomic DNA was extracted from stable cell lines and various different fragments of the expression cassette amplified by PCR using these primers and their

size analysed on an agarose gel. Untransduced A549 cells were used as a negative control and SET7-Clone3 serves as a positive control in this PCR assay. Fragments of the expected lengths could be amplified for SET7-Clone3, Clone4 and Clone12 (**Figure 5.2**, lanes D, E, and F respectively) suggesting that these cell lines harbour a complete expression cassette integrated into the genome as expected. To my surprise, I found that SET7-Clone1 and all three p300 clones have lost parts of the promoter and the dCas9-chromatin modifier construct and have only kept the Blasticidin resistance gene (**Figure 5.2**, primers Blast-9F and Blast-9R).

I also extracted RNA from the monoclonal stable cell lines. qPCR reactions to test for the presence of dCas9-p300 and dCas9-SET7 mRNA in these samples were performed by Patricia deWinter at qStandard using normalisation against two reference genes (*GUSB* and *YWHAZ*). In line with my results shown in **Figure 5.2**, there was evidence for mRNA expression of the dCas9-chromatin modifier construct in SET7-Clones3 and SET7-Clone12, but not for any of the other SET7 or p300 monoclonal stable cell lines.

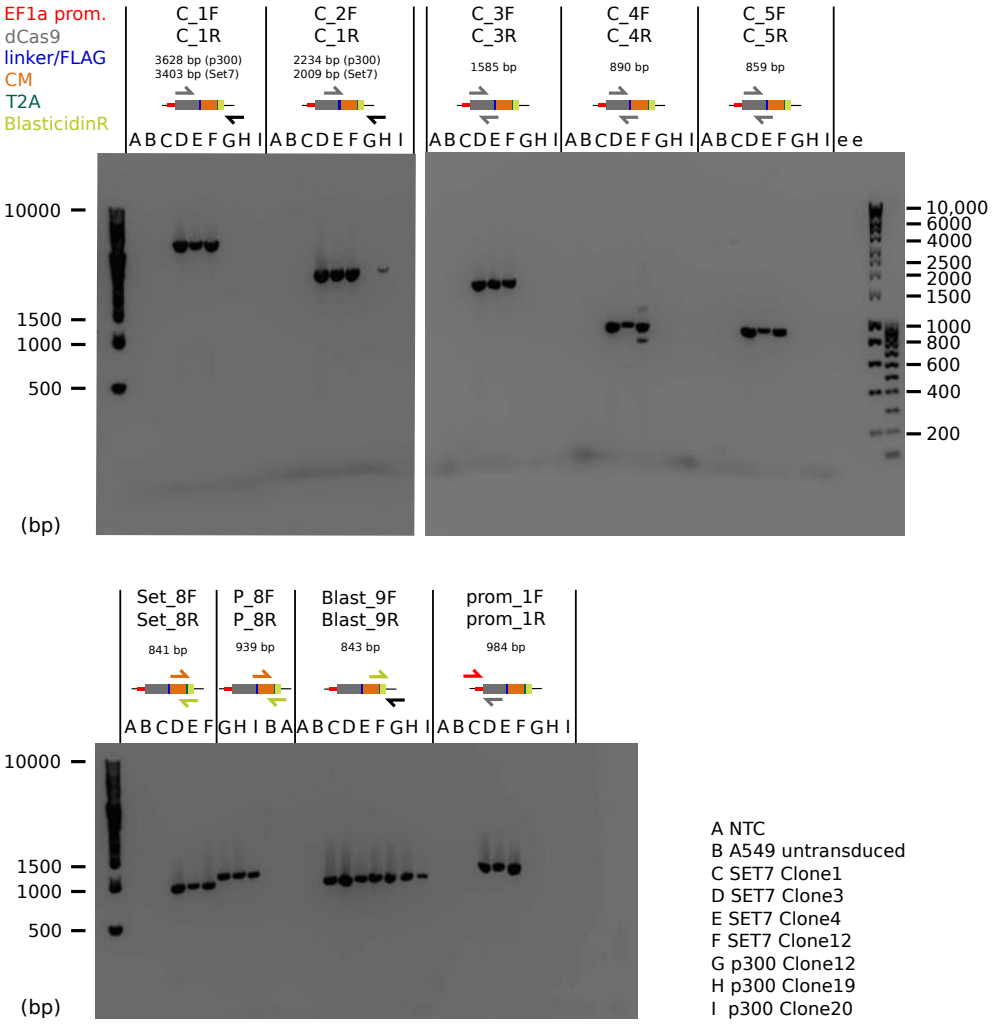


Figure 5.2: Amplification of dCas9 constructs from genomic DNA extracted from monoclonal cell lines. Different parts of the dCas9-chromatin modifier-T2A-Blasticidin expression cassette were amplified from genomic DNA extracted from monoclonal stable cell lines using the primer pairs indicated and analysed on a 1 % agarose gel stained with GelRed (Biotium). The 1kb DNA ladder (NEB), Hyper Ladder I and Hyper Ladder IV (both Bioline) were run for comparison. NTC denotes the no-template control.

Given these results, I next asked whether the expression cassettes can be detected in the original polyclonal cell lines (see **Figure 3.6**). Using the same PCR primers and conditions and including the monoclonal SET7 Clone3 cells as a positive control, I confirmed the presence of full-length expression cassettes in the original polyclonal pools (**Figure 5.3**). This suggests it should be possible to return to the original low-passage polyclonal pool of cells to establish new monoclonal stable cell lines.

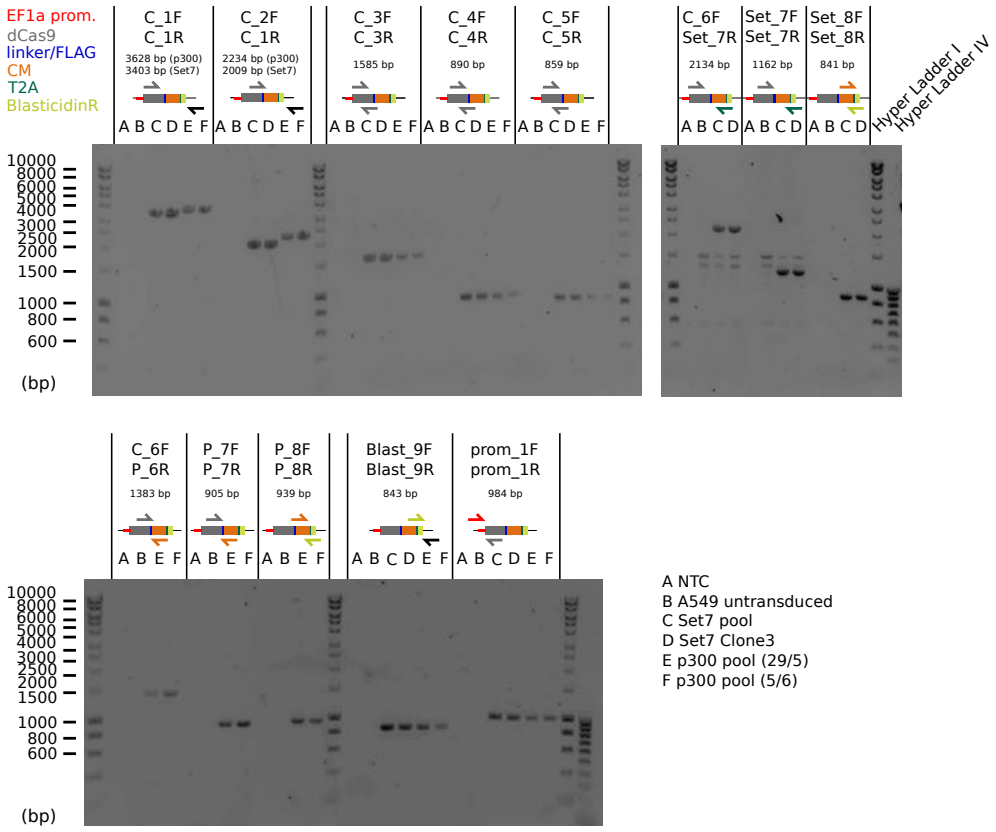


Figure 5.3: Amplification of dCas9 constructs from genomic DNA extracted from polyclonal cell lines. Different parts of the dCas9-chromatin modifier-T2A-Blasticidin expression cassette were amplified from genomic DNA extracted from polyclonal stable cell lines (labelled "pools") using the primer pairs indicated and analysed on a 1 % agarose gel stained with Sybr Safe (Invitrogen). Hyper Ladder I and Hyper Ladder IV (both Bioline) were run for comparison. NTC denotes the no-template control. gDNA from the monoclonal cell line SET7-Clone3 was included as a positive control.

5.3 Discussion

In this chapter, I described the validation experiments to test whether any of the candidate gRNAs identified from the screens using the EMT5000 gRNA library and A549 cells expressing the dCas9-SET7 construct could induce cadherin switching when introduced into these cells individually. When all candidate gRNAs failed validation, I decided to first check expression of the dCas9-chromatin modifier constructs in the cell lines I had selected for the screening experiments. Western blots had previously only confirmed expression of the dCas9-SET7 chromatin modifier in the SET7-Clone3 line and I had always assumed this was due in part to the sensitivity of the assay. I harvested both DNA and RNA from

the different monoclonal stable cell lines in order to assess both integration of the chromatin modifier expression cassette into the genome and expression of the chromatin modifier at the level of mRNA. To my surprise, I found that some of the cell lines (SET7-Clone1 and p300-Clone12, Clone19 and Clone20) had lost large parts of the construct. SET7-Clone1 has lost the EF1a promoter, dCas9 and the SET7 sequence. All three dCas9-p300 cell lines appear to have lost the EF1a promoter sequence and all of dCas9, but have kept parts of p300. It is unclear whether loss of these sequences occurred upon integration of the cassette after viral infection or whether the sequence was excised later. All cell lines still harbored the Blasticidin resistance gene sequence, which is consistent with the observation that these cells were resistant to Blasticidin. The most likely explanation for how an integrated Blasticidin gene without a promoter or ribosome entry site can still confer Blasticidin resistance to the cells is that the partial cassette might have integrated in-frame and in the correct orientation into an actively transcribed gene downstream of an endogenous promoter (**Figure 5.4**). This is consistent with the observation that lentiviral integration predominantly occurs in actively transcribed regions of the genome [155].

I then wondered whether any cells with a correctly inserted expression cassette could be detected in the original polyclonal pool from which the monoclonal stable cell lines had been established. I was able to confirm the presence of full-length expression cassettes in DNA extracted from the original pools, which means that it should be possible to establish new monoclonal stable cell lines for future screening experiments from these. I previously used immunofluorescence staining to identify those clones that expressed dCas9 (section 3.1.3). While brightly staining cells could still be identified in the polyclonal cell lines, these were no longer seen following single cell dilution and many rounds of division to establish monoclonal lines. It appears that many cells that expressed relatively high levels of the targetable chromatin modifier did not survive single cell dilution and continued passage. It is also possible that the cells lost the integrated expression cassette through excision or a recombination event.

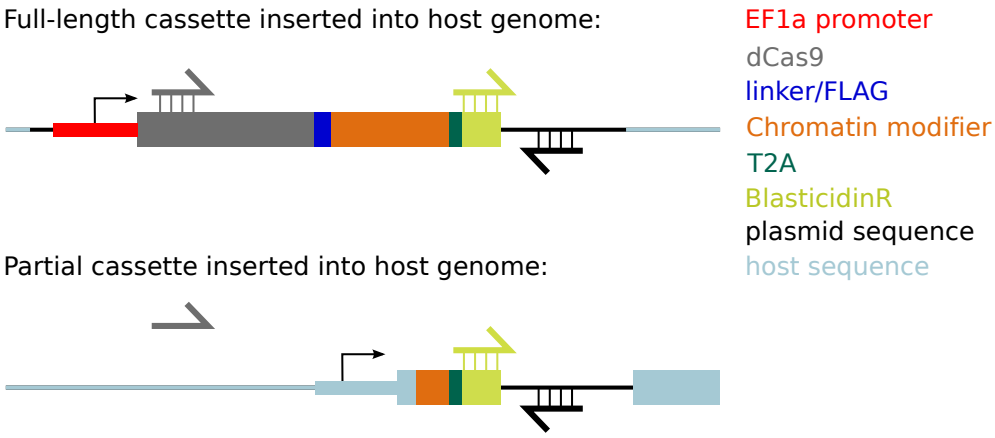


Figure 5.4: Integration of the dCas9-chromatin modifier expression cassette into the A549 genome.

Given that cells harbouring a complete expression cassette may be relatively rare in the population (see **Figure 3.6**), I would increase the number of clones to be screened following single-cell dilution in the future. I would further use a PCR-based assay (with primers C-1F and C-1R as shown in **Figure 5.2**) as a pre-selection followed by Western blotting rather than immunofluorescence staining to identify clonal lines that express the targetable chromatin modifier.

Chapter 6

Discussion

6.1 Discussion and future directions

6.1.1 Chromatin modifiers

It is clear that expression of the targetable chromatin modifier is a pre-requisite for a successful epigenetic screen. After none of the candidate gRNAs I identified in screens with the dCas9-p300 and dCas9-SET7 chromatin modifier could be experimentally validated, I discovered that many of the monoclonal cell lines (with the exception of three of the SET7 clonal lines) harbour genetic deletions of the chromatin modifier expression cassette. Only in SET7-Clone3 could a band of the size of the targetable chromatin modifier be reliably detected by Western blotting using an anti FLAG antibody.

Although the sequences to be packaged into lentivirus are large (10 kb for dCas9-p300, 9.6 kb for dCas9-SET7), their length is still below the upper size limit for lentivirus production [156]. Of note, loss of large parts of the construct would not necessarily be reflected in the virus titer as long as the Blasticidin resistance gene is still successfully expressed following integration into the host genome. It is possible that loss of parts of the expression cassette (as seen in all three dCas9-p300 clones and in SET7-Clone1) occurred during lentivirus packaging or upon integration into the host genome. Alternatively, excision from the genome could have occurred at some point after integration.

As mentioned in section 5.3, it should be possible to return to the original polyclonal stable cell lines I established and derive new clones that harbour a complete expression cassette. In hindsight, I should have perhaps used a lentiviral dCas9-chromatin modifier-T2A-GFP fusion construct instead of the dCas9-chromatin modifier-T2A-Blasticidin sequence. This would have facilitated establishment of monoclonal stable cell lines as GFP-positive single cells could have been isolated by FACS. However, the GFP sequence could integrate and be successfully expressed without the upstream part of the expression cassette just as the Blasticidin sequence.

While the problems I encountered with establishing monoclonal cell lines expressing the targetable chromatin modifiers certainly need to be solved in order to make epigenetics screens a reality, I have also identified several potential stumbling blocks in the screening protocol I employed, some of which I was able to overcome while others still remain to be addressed in the future as discussed below. The steps of the screening protocol are shown in **Figure 6.1** in order to

illustrate some of the critical points that could be optimised.

6.1.2 Monitoring gRNA library representation

Library complexity can be assessed by next-generation sequencing once the gRNA library has been made. However, for very complex libraries this is costly. My collaborators estimated the complexity of the MNase-digest libraries from sequencing data (section 3.2.2) but as a quality-control step and for troubleshooting purposes, it would have been useful to keep track of the library representation during library preparation. I performed control electroporations using a pUC19 plasmid to ensure the home-made electro-competent TG1 bacteria I used for library preparation have a competency of $> 10^{10}$ colony-forming units per μg DNA. However, I should have also plated a dilution series of the same bacteria transformed with the library so as to be able to count the number of colonies per μg library DNA. This would have ensured that library complexity and representation is maintained from cloning of gRNAs into the expression vector to harvesting of the plasmid DNA. Following packaging of the gRNA library into lentivirus it is also important to maintain sufficient representation (around 20-100 fold coverage) by ensuring enough cells are successfully transduced (see section 4.3).

6.1.3 Bias introduced by FACS sorting

Following transduction of cells with the lentiviral gRNA library, cells of interest are isolated from the population by FACS. FACS constitutes a strong selection that in my case resulted in low numbers of cells being selected for downstream analysis. FACS further introduces false-positives. Furthermore, I was isolating cells that had lost expression of the cell surface marker E-cadherin. Incomplete staining by the E-cadherin antibody can thus also contribute to false-positive signal. This introduces random noise into the data that makes it more difficult to identify reliable candidates later on. When I first came up with the strategy for the CRISPR-based epigenetic screening method, no CRISPR-based screens had been published. While a range of genetic and activator/inhibitor screens have now been reported, only a single study published to date has used FACS sorting to enrich for cells of interest in a CRISPR-based screen [100]. This may reflect the variability introduced by using FACS which leads to difficulties in the downstream analysis that I was originally not aware of.

While I pursued a strategy aimed at minimizing the selection of false-positive cells by using very stringent gates for sorting (selecting only 1-5 % of the population), it may in future be better to sort less stringently. In addition, the number of cells used for sorting could be scaled up. Together, this would greatly increase the number of sorted cells available for downstream analysis. This would in turn permit the use of fewer cycles of amplification before sequencing which would also reduce PCR bias. When choosing gates so as to sort perhaps 15-20 % of the total population, a lot of false-positive cells will be sorted but these are selected at random and, given sufficient numbers of sorted cells, will cover the entire library. Programmes such as DESeq2, written for identification of differentially expressed genes from RNA-seq data could be again used to identify those gRNAs that are enriched ("over-expressed") in the samples relative to the negative controls.

6.1.4 PCR amplification bias and sequencing depth

PCR amplification bias is a problem in any next-generation sequencing work-flow, but is usually ignored. However, when the aim of the analysis is to accurately count and compare certain sequences in samples relative to controls, it is particularly important to consider the effect of this bias. PCR amplification bias is random and leads to some sequences becoming over-represented and other sequences under-represented after amplification. Whether a sequence becomes preferentially amplified during PCR may depend on its GC content, length and conformation of the DNA molecule. Most amplicons I sequenced differ only in the central 20 bp that correspond the gRNA protospacer sequence, while the surrounding plasmid sequences are identical for all molecules in the library. This means that the length of all amplicons are identical and the overall GC content is likely very similar for all molecules in the library, making it unlikely that these two factors could generate large bias. However, simulations suggest that whether a sequence is preferentially amplified may even depend on factors such as position of molecule within the reaction vessel [148]. In addition, there is a second stochastic process that may introduce variation into the data, namely random sampling from the amplified product by the sequencing process.

I employed a strategy that involved addition of unique molecular identifiers (UMIs) to gRNA sequences before sequencing library preparation (see section 4.3.1). This strategy has been established for single cell RNA-seq protocols and can be used to infer the number of template molecules prior to amplification and sequencing [149, 150]. However, UMIs are usually added to sequences of interest

during reverse transcription (for RNA-seq) or by ligation. Because the gRNA sequence was integrated into the genome I needed to incorporate addition of UMIs into an amplicon sequencing approach and therefore decided to add UMIs to the gRNA sequences by an initial round of PCR amplification.

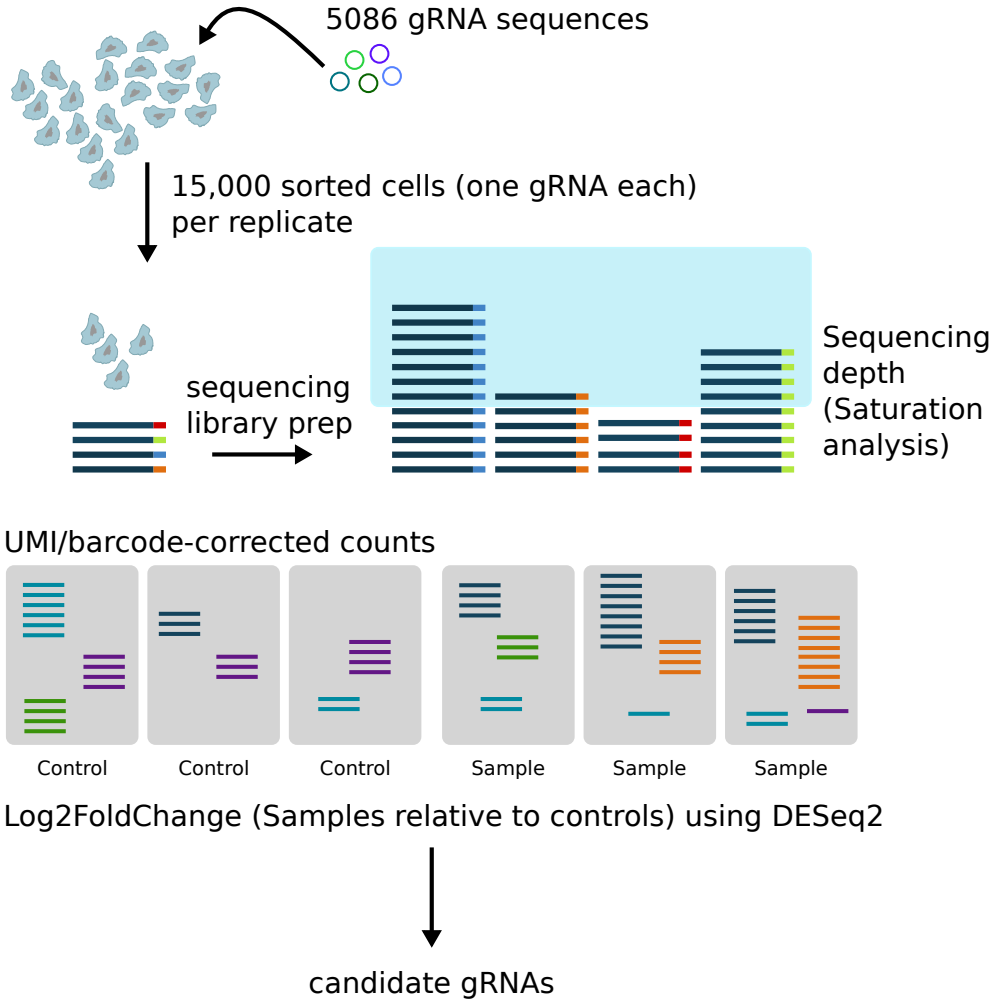


Figure 6.1: Schematic illustration of steps in the screening protocol.

Incorporation of UMIs prior to amplification allowed me to correct for some of the PCR amplification bias. However, this strategy means that a low amount of starting material is massively amplified which then necessitates relatively deep sequencing to sample enough of the population of molecules in which some molecules have become disproportionately enriched (see **Figure 6.1**). In future, it may become possible to sequence a low amount of starting material directly with single molecule amplification-free DNA sequencing methods, which are currently under development. For now, increasing the amount of starting material would make it possible to reduce the number of required cycles of PCR amplification, which would in turn reduce bias. However, this is difficult to achieve in my case, given that FACS substantially limits the amount of starting material available for sequencing.

6.1.5 EMT as a phenotypic readout

It should be possible to test whether the current screening protocol is able to identify signal if it exists. One option would be to generate E-Cadherin knockout cells. These cells could then be infected with a single gRNA that is not present in the library, e.g. the control gRNA against *GAPDH* that was used in the validation experiments. Given that these cells can no longer express E-Cadherin, the *GAPDH* gRNA should be identified as a candidate by the downstream analysis. The cells expressing the *GAPDH* gRNA but no E-Cadherin protein could then be mixed at different ratios with wild-type cells transduced with the EMT5000 library. The different mixtures of cells would then be subjected to FACS sorting and gRNA sequences counted by next-generation sequencing as before. By using different amounts of the spike-in of the "pseudo-functional" *GAPDH* gRNA it should be possible to determine the sensitivity of the current screening protocol to detect functional gRNAs.

It is also possible that variability in antibody staining contributes to the observed high levels of variability in the data. The need for antibody staining could be eliminated entirely by making cell lines with E-Cadherin and N-Cadherin genes endogenously tagged with fluorescent proteins. One caveat of this strategy is the need for the GFP fusion protein to be folded and exported to cell surface correctly. This would have to be established first. In FosER cells, an exogenous E-Cadherin-GFP fusion protein has previously been reported to localise correctly to the plasma membrane when expressed at low levels [157]. It may also be possible to reduce the rate of false-positives by culturing E-Cadherin negative

cells after sorting, followed by re-staining and re-sorting.

Another option to circumvented many of the problems discussed above would be to identify a method other than FACS to isolate cells that have undergone an EMT-like change in phenotype. For example, trans-well migration assays have previously been performed on A549 cells [158]. However, these migration assays also suffer from false-positives, i.e. some cells are able to migrate through the membrane prior to EMT induction already. Throughput may also be an issue with these types of assays.

6.1.6 Switching to a different experimental system

Currently, the key challenge with using cadherin switching (or EMT) as a phenotypic readout for an epigenetic screen is the lack of a positive control, i.e. a gRNA that is known *a priori* to induce an EMT-like phenotype in cells when co-expressed with a particular chromatin modifier. By performing a small-scale screen using the EMT5000 library I was hoping to identify such a functional gRNA that could then be used in more unbiased screens with the MNase digest library. At the time the only publications using targetable chromatin modifiers looked for changes in transcript levels [75, 77], with few exceptions where changes in protein levels were also measured following epigenome engineering [79]. Changes at the level of cellular behaviour or phenotype have only been examined more recently both *in vivo* [84, 87] and in *in vitro* models [79, 87] (see also **Tables 1.1** and **1.2**).

Several examples have now been reported in the literature where epigenome engineering results in changes at the level of protein or phenotype, some of which could be suitable systems for establishing an epigenetic screening method given that the reports provide a positive control gRNA. For example Rivenbark *et al.* reported downregulation of the Maspin protein, encoded by the *SERPINB5* gene, following targeted DNA methylation with a DNMT3A construct [79]. This was accompanied by a decrease in protein level and increased ability to form colonies in SUM159 breast cancer cells. While it would be difficult to use the colony-formation assay to isolate cells of interest given the observed rate of false-positives, this system could nevertheless be adapted for use in a screen, for example by endogenously tagging the MASPIN protein with a fluorescent protein. Similarly, targeted DNA demethylation at the promoter of the *MYOD* gene in C3H10T1/2 cells using a dCas9-TET1 construct led to measurable increase in protein levels

[87]. This system could again be adapted for screening by endogenous tagging with a fluorescent protein.

6.1.7 Genome-wide screens with ultra-complex gRNA libraries

The aim of screen with EMT5000 control library was to identify a gRNA that together with a targetable chromatin modifier could induce loss of E-Cadherin in A549 cells. This gRNA could have then served as a positive control for an unbiased screen using the much more complex genome-wide libraries to discover new sites in the genome that can be activated or silenced by chromatin modifiers and have an impact on EMT phenotype. In this PhD thesis, I developed a novel method to generate ultra-complex gRNA libraries from any source of DNA using enzymatic digestion by micrococcal nuclease (see **Figure 3.10** and section 2.2.3). This method could be of general use both for genetic and epigenetic CRISPR-based screens. While the presence of a PAM sequence downstream of the gRNA targeting site is not guaranteed (compared to a similar method [122]) the method using micrococcal nuclease generates all possible gRNAs from a particular region and could be adapted for generating gRNA libraries not just from purified genomic DNA but also from PCR amplicons or BAC clones of regions of interest or from immunoprecipitated DNA.

6.2 Conclusion

The aim of this PhD thesis was to develop a CRISPR-based screening method to identify sites in the genome where the activity of a chromatin-modifying enzyme can induce a change in cellular phenotype. I successfully developed a complete work-flow for this, including generation of the required constructs and cell lines as well as experimental and computational procedures: (1) I generated eight different targetable chromatin modifier constructs that are able to add or remove different chromatin marks from both histones and DNA bases. (2) I used lentiviral transduction to establish stable cell lines expressing these constructs. (3) I also developed a novel method for generating gRNA libraries based on digestion of genomic DNA with micrococcal nuclease. These libraries have unprecedented sequence complexity and may in future be used in large-scale unbiased genetic and epigenetic screens. (4) I further established cadherin switching, which occurs

during epithelial-to-mesenchymal transition (EMT), as a potential phenotypic read-out for my epigenetic screening method. (5) I then conducted the first epigenetic CRISPR-based screen using a library containing 5000 gRNAs that target the promoters of 15 genes known to be involved in the regulation of EMT together with the cell lines expressing the dCas9-p300 histone acetyltransferase or dCas9-SET7 histone methyltransferase constructs. FACS was used to identify and isolate cells of interest that displayed loss of E-cadherin expression following epigenome engineering. I then amplified the gRNA targeting sequences from these cells for next-generation sequencing. (6) I performed data analysis to identify candidate gRNAs that had become enriched in the cells that had lost E-Cadherin. Together with my collaborator, I developed computational methods for accurate counting of gRNAs. (7) I subsequently attempted to experimentally validate the candidates I had identified. As none of the candidate gRNAs could be experimentally validated, I outlined several possible routes for future optimization of the screening method. As such, this work constitutes an important first step towards establishing epigenome-wide screens to identify sites in the genome where chromatin modification really matters.

Bibliography

[1] R. Jaenisch and A. Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33 (3s):245–254, 2003.

[2] C. H. Waddington. The Epigenotype. *International Journal of Epidemiology*, 41(1):10–13, 2012.

[3] D. F. Browning and S. J. W. Busby. The regulation of bacterial transcription initiation. *Nature Reviews Microbiology*, 2(1):57–65, 2004.

[4] C. Wu. Heat shock transcription factors: structure and regulation. *Annual Review of Cell and Developmental Biology*, 11:441–469, 1995.

[5] A. Hoffmann and D. Baltimore. Circuitry of nuclear factor kappaB signaling. *Immunological reviews*, 210:171–186, 2006.

[6] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Publishing Group*, 10(4):252–263, 2009.

[7] A. Jolma, Y. Yin, K. R. Nitta, K. Dave, A. Popov, M. Taipale, M. Enge, T. Kivioja, E. Morgunova, and J. Taipale. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, 527(7578): 384–388, 2015.

[8] Y. Cao, Z. Yao, D. Sarkar, M. Lawrence, G. J. Sanchez, M. H. Parker, K. L. MacQuarrie, J. Davison, M. T. Morgan, W. L. Ruzzo, R. C. Gentleman, and S. J. Tapscott. Genome-wide MyoD Binding in Skeletal Muscle Cells: A Potential for Broad Cellular Reprogramming. *Developmental Cell*, 18(4): 662–674, 2010.

[9] Y. C. Lin, S. Jhunjhunwala, C. Benner, S. Heinz, E. Welinder, R. Mansson, M. Sigvardsson, J. Hagman, C. A. Espinoza, J. Dutkowski, T. Ideker, C. K. Glass, and C. Murre. A global network of transcription factors, in-

volving E2A, EBF1 and Foxo1, that orchestrates the B cell fate. *Nature Immunology*, 11(7):635–643, 2010.

[10] A. M. Pilon, S. S. Ajay, S. A. Kumar, L. A. Steiner, P. F. Cherukuri, S. Wincovitch, S. M. Anderson, NISC Comparative Sequencing Center, J. C. Mullikin, P. G. Gallagher, R. C. Hardison, E. H. Margulies, and D. M. Bodine. Genome-wide ChIP-Seq reveals a dramatic shift in the binding of the transcription factor erythroid Kruppel-like factor during erythrocyte differentiation. *Blood*, 118(17):e139–e148, 2011.

[11] F. Gao, B. C. Foat, and H. J. Bussemaker. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, 5:31, 2004.

[12] S. A. Vokes, H. Ji, W. H. Wong, and A. P. McMahon. A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb. *Genes & Development*, 22(19):2651–2663, 2008.

[13] C. R. Lickwar, F. Mueller, S. E. Hanlon, J. G. McNally, and J. D. Lieb. Genome-wide protein–DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature*, 484(7393):251–255, 2012.

[14] P. Tessarz and T. Kouzarides. Histone core modifications regulating nucleosome structure and dynamics. *Nature Reviews Molecular Cell Biology*, 15 (11):703–708, 2014.

[15] T. J. Richmond and C. A. Davey. The structure of DNA in the nucleosome core. *Nature*, 423(6936):145–150, 2003.

[16] B. D. Strahl and C. D. Allis. The language of covalent histone modifications. *Nature*, 403(6765):41–45, 2000.

[17] P. A. Wade, D. Pruss, and A. P. Wolffe. Histone acetylation: chromatin in action. *Trends in Biochemical Sciences*, 22(4):128–132, 1997.

[18] S. L. Schreiber and B. E. Bernstein. Signaling network model of chromatin. *Cell*, 111(6):771–778, 2002.

[19] A. J. Bannister, P. Zegerman, J. F. Partridge, E. A. Miska, J. O. Thomas, R. C. Allshire, and T. Kouzarides. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature*, 410(6824):120–124, 2001.

[20] M. Lachner, D. O'Carroll, S. Rea, K. Mechtler, and T. Jenuwein. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature*, 410(6824):116–120, 2001.

[21] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837, 2007.

[22] Z. Wang, C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, W. Peng, M. Q. Zhang, and K. Zhao. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 40(7):897–903, 2008.

[23] H. Santos-Rosa, R. Schneider, A. J. Bannister, J. Sherriff, B. E. Bernstein, N. C. T. Emre, S. L. Schreiber, J. Mellor, and T. Kouzarides. Active genes are tri-methylated at K4 of histone H3. *Nature*, 419(6905):407–411, 2002.

[24] B. E. Bernstein, M. Kamal, K. Lindblad-Toh, S. Bekiranov, D. K. Bailey, D. J. Huebert, S. McMahon, E. K. Karlsson, E. J. Kulbokas, III, T. R. Gingeras, S. L. Schreiber, and E. S. Lander. Genomic Maps and Comparative Analysis of Histone Modifications in Human and Mouse. *Cell*, 120(2): 169–181, 2005.

[25] T. H. Kim, L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, T. A. Richmond, Y. Wu, R. D. Green, and B. Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880, 2005.

[26] C.-K. Lee, Y. Shibata, B. Rao, B. D. Strahl, and J. D. Lieb. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics*, 36(8):900–905, 2004.

[27] R. Karlic, H. R. Chung, J. Lasserre, K. Vlahovicek, and M. Vingron. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, 107(7):2926–2931, 2010.

[28] T. K. Barth and A. Imhof. Fast signals and slow marks: the dynamics of histone modifications. *Trends in Biochemical Sciences*, 35(11):618–626, 2010.

[29] `http://ihec-epigenomes.org/`.

[30] `https://www.encodeproject.org/`.

[31] `http://www.roadmapepigenomics.org`.

[32] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 488(7414):57–74, 2012.

[33] F. J. Iborra, A. Pombo, D. A. Jackson, and P. R. Cook. Active RNA polymerases are localized within discrete transcription "factories" in human nuclei. *Journal of Cell Science*, 109 (Pt 6):1427–1436, 1996.

[34] A. Bolzer, G. Kreth, I. Solovei, D. Koehler, K. Saracoglu, C. Fauth, S. Müller, R. Eils, C. Cremer, M. R. Speicher, and T. Cremer. Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. *PLoS Biology*, 3(5):e157, 2005.

[35] M. V. Rudan, C. Barrington, S. Henderson, C. Ernst, D. T. Odom, A. Tanay, and S. Hadjur. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Reports*, 10(8): 1297–1309, 2015.

[36] C. Kimura-Yoshida, K. Kitajima, I. Oda-Ishii, E. Tian, M. Suzuki, M. Yamamoto, T. Suzuki, M. Kobayashi, S. Aizawa, and I. Matsuo. Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development*, 131(1): 57–71, 2004.

[37] M. H. Kagey, J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. van Berkum, C. C. Ebmeier, J. Goossens, P. B. Rahl, S. S. Levine, D. J. Taatjes, J. Dekker, and R. A. Young. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435, 2010.

[38] B. Tolhuis, R.-J. Palstra, E. Splinter, F. Grosveld, and W. de Laat. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell*, 10(6):1453–1465, 2002.

[39] Y. Guo, Q. Xu, D. Canzio, J. Shou, J. Li, D. U. Gorkin, I. Jung, H. Wu, Y. Zhai, Y. Tang, Y. Lu, Y. Wu, Z. Jia, W. Li, M. Q. Zhang, B. Ren, A. R. Krainer, T. Maniatis, and Q. Wu. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell*, 162(4): 900–910, 2015.

[40] C. G. Spilianakis, M. D. Lalioti, T. Town, G. R. Lee, and R. A. Flavell. Interchromosomal associations between alternatively expressed loci. *Nature*, 435(7042):637–645, 2005.

[41] A. Papantonis, J. D. Larkin, Y. Wada, Y. Ohta, S. Ihara, T. Kodama,

and P. R. Cook. Active RNA Polymerases: Mobile or Immobile Molecular Machines? *PLoS Biology*, 8(7):e1000419, 2010.

[42] S. Fanucchi, Y. Shibayama, S. Burd, M. S. Weinberg, and M. M. Mhlanga. Chromosomal Contact Permits Transcription between Coregulated Genes. *Cell*, 155(3):606–620, 2013.

[43] C. S. Osborne, L. Chakalova, K. E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J. A. Mitchell, S. Lopes, W. Reik, and P. Fraser. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature Genetics*, 36(10):1065–1071, 2004.

[44] B. Tolhuis, R.-J. Palstra, E. Splinter, F. Grosveld, and W. de Laat. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell*, 10(6):1453–1465, 2002.

[45] J. M. Brown, J. Leach, J. E. Reittie, A. Atzberger, J. Lee-Prudhoe, W. G. Wood, D. R. Higgs, F. J. Iborra, and V. J. Buckle. Coregulated human globin genes are frequently in spatial proximity when active. *The Journal of cell biology*, 172(2):177–187, 2006.

[46] L. A. Cirillo, F. R. Lin, I. Cuesta, D. Friedman, M. Jarnik, and K. S. Zaret. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Molecular Cell*, 9(2):279–289, 2002.

[47] M. Lupien, J. Eeckhoute, C. A. Meyer, Q. Wang, Y. Zhang, W. Li, J. S. Carroll, X. S. Liu, and M. Brown. FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, 132(6):958–970, 2008.

[48] S. Pérez-Lluch, E. Blanco, H. Tilgner, J. Curado, M. Ruiz-Romero, M. Corominas, and R. Guigo. Absence of canonical marks of active chromatin in developmentally regulated genes. *Nature Genetics*, 47(10):1158–1167, 2015.

[49] X. Dong, M. C. Greven, A. Kundaje, S. Djebali, J. B. Brown, C. Cheng, T. R. Gingeras, M. Gerstein, R. Guigo, E. Birney, and Z. Weng. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology*, 13(9):R53, 2012.

[50] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham,

M. Kellis, and B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.

[51] B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, and E. S. Lander. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, 125(2):315–326, 2006.

[52] A. Wutz, O. W. Smrzka, N. Schweifer, K. Schellander, E. F. Wagner, and D. P. Barlow. Imprinted expression of the Igf2r gene depends on an intronic CpG island. *Nature*, 389(6652):745–749, 1997.

[53] A. Murrell, S. Heeson, and W. Reik. Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nature Genetics*, 36(8):889–893, 2004.

[54] P. A. Latos, F. M. Pauler, M. V. Koerner, H. B. Şenergin, Q. J. Hudson, R. R. Stocsits, W. Allhoff, S. H. Stricker, R. M. Klement, K. E. Warczok, K. Aumayr, P. Pasierbek, and D. P. Barlow. Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science*, 338(6113):1469–1472, 2012.

[55] E. Li, T. H. Bestor, and R. Jaenisch. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–926, 1992.

[56] M. Okano, D. W. Bell, D. A. Haber, and E. Li. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.

[57] J. Wang, K. Scully, X. Zhu, L. Cai, J. Zhang, G. G. Prefontaine, A. Krones, K. A. Ohgi, P. Zhu, I. Garcia-Bassets, F. Liu, H. Taylor, J. Lozach, F. L. Jayes, K. S. Korach, C. K. Glass, X.-D. Fu, and M. G. Rosenfeld. Opposing LSD1 complexes function in developmental gene activation and repression programmes. *Nature*, 446(7138):882–887, 2007.

[58] G. Lagger, D. O'Carroll, M. Rembold, H. Khier, J. Tischler, G. Weitzer, B. Schuettengruber, C. Hauser, R. Brunmeir, T. Jenuwein, and C. Seiser. Essential function of histone deacetylase 1 in proliferation control and CDK inhibitor repression. *The EMBO Journal*, 21(11):2672–2681, 2002.

[59] A. H. Peters, D. O'Carroll, H. Scherthan, K. Mechtler, S. Sauer, C. Schöfer,

K. Weipoltshammer, M. Pagani, M. Lachner, A. Kohlmaier, S. Opravil, M. Doyle, M. Sibilia, and T. Jenuwein. Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell*, 107(3):323–337, 2001.

[60] C. L. Smith and J. T. Eppig. Expanding the mammalian phenotype ontology to support automated exchange of high throughput mouse phenotyping data generated by large-scale mouse knockout screens. *Journal of Biomedical Semantics*, 6(11):1–7, 2015.

[61] K. Funato, T. Major, P. W. Lewis, C. D. Allis, and V. Tabar. Use of human embryonic stem cells to model pediatric gliomas with H3.3K27M histone mutation. *Science*, 346(6216):1529–1533, 2014.

[62] V. Boonsanay, T. Zhang, A. Georgieva, S. Kostin, H. Qi, X. Yuan, Y. Zhou, and T. Braun. Regulation of Skeletal Muscle Stem Cell Quiescence by Suv4-20h1-Dependent Facultative Heterochromatin Formation. *Stem Cell*, 18(2): 229–242, 2016.

[63] R. Hamamoto, V. Saloura, and Y. Nakamura. Critical roles of non-histone protein lysine methylation in human tumorigenesis. *Nat. Rev. Cancer*, 15 (2):110–124, 2015.

[64] A. J. Bannister, E. A. Miska, D. Görlich, and T. Kouzarides. Acetylation of importin-alpha nuclear import factors by CBP/p300. *Current Biology*, 10(8):467–470, 2000.

[65] W. Gu and R. G. Roeder. Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain. *Cell*, 90(4):595–606, 1997.

[66] S. Domcke, A. F. Bardet, P. A. Ginno, D. Hartl, L. Burger, and D. Schübeler. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature*, 528(7583):575–579, 2015.

[67] N. Kanu, E. Grönroos, P. Martinez, R. A. Burrell, X. Y. Goh, J. Bartkova, A. Maya-Mendoza, M. M. i. k, A. J. Rowan, H. Patel, A. Rabinowitz, P. East, G. Wilson, C. R. Santos, N. McGranahan, S. Gulati, M. Gerlinger, N. J. Birkbak, T. Joshi, L. B. Alexandrov, M. R. Stratton, T. Powles, N. Matthews, P. A. Bates, A. Stewart, Z. Szallasi, J. Larkin, J. Bartek, and C. Swanton. SETD2 loss-of-function promotes renal cancer branched evolution through replication stress and impaired DNA repair. *Oncogene*, 34(46):5699–5708, 2015.

[68] T. Gaj, C. A. Gersbach, and C. F. Barbas, III. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology*, 31(7):397–405, 2013.

[69] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, 337(6096):816–821, 2012.

[70] P. Mali, L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, and G. M. Church. RNA-Guided Human Genome Engineering via Cas9. *Science*, 339(6121):823–826, 2013.

[71] L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*, 339(6121):819–823, 2013.

[72] K. M. Esvelt, P. Mali, J. L. Braff, M. Moosburner, S. J. Yaung, and G. M. Church. Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nature Methods*, 10:1116–1121, 2013.

[73] S. H. Stricker, A. Köferle, and S. Beck. From profiles to function in epigenomics. *Nature Reviews Genetics*, 18(1):51–66, 2017.

[74] I. B. Hilton, A. M. D'Ippolito, C. M. Vockley, P. I. Thakore, G. E. Crawford, T. E. Reddy, and C. A. Gersbach. Epigenome Editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature Biotechnology*, 33:1–10, 2015.

[75] M. L. Maeder, J. F. Angstman, M. E. Richardson, S. J. Linder, V. M. Cascio, S. Q. Tsai, Q. H. Ho, J. D. Sander, D. Reyon, B. E. Bernstein, J. F. Costello, M. F. Wilkinson, and J. K. Joung. Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins . *Nature Biotechnology*, 31(12):1137–1142, 2013.

[76] D. Cano-Rodriguez, R. A. F. Gjaltema, L. J. Jilderda, P. Jellema, J. Dokter-Fokkens, M. H. J. Ruiters, and M. G. Rots. Writing of H3K4Me3 overcomes epigenetic silencing in a sustained but context-dependent manner. *Nature Communications*, 7:1–11, 2016.

[77] E. M. Mendenhall, K. E. Williamson, D. Reyon, J. Y. Zou, O. Ram, J. K. Joung, and B. E. Bernstein. Locus-specific editing of histone modifications at endogenous enhancers. *Nature Biotechnology*, 31(12):1133–1136, 2013.

[78] N. A. Kearns, H. Pham, B. Tabak, R. M. Genga, N. J. Silverstein, M. Gar-

ber, and R. e. Maehr. Functional Annotation of native enhancers with a Cas9-histone demethylase fusion. *Nature Methods*, 12(5):401–403, 2015.

[79] A. G. Rivenbark, D. J. McKay, S. Stolzenburg, J. D. Lieb, A. S. Beltran, X. Yuan, M. G. Rots, B. D. Strahl, and P. Blancafort. Epigenetic reprogramming of cancer cells via targeted DNA methylation. *Epigenetics*, 7(4): 350–360, 2012.

[80] A. N. Siddique, S. Nunna, A. Rajavelu, Y. Zhang, R. Z. Jurkowska, R. Reinhardt, M. G. Rots, S. Ragozin, T. P. Jurkowski, and A. Jeltsch. Targeted methylation and gene silencing of VEGF-A in human cells by using a designed Dnmt3a-Dnmt3L single-chain fusion protein with increased DNA methylation activity. *Journal of Molecular Biology*, 425(3):479–491, 2013.

[81] D. L. Bernstein, J. E. Le Lay, E. G. Ruano, and K. H. Kaestner. TALE-mediated epigenetic suppression of CDKN2A increases replication in human fibroblasts. *Journal of Clinical Investigation*, 125(5):1998–2006, 2015.

[82] A. Vojta, P. Dobrinić, V. Tadić, L. Bočkor, P. Korać, B. Julg, M. Klasić, and V. Zoldoš. Repurposing the CRISPR-Cas9 system for targeted DNA methylation. *Nucleic Acids Research*, pages 1–14, 2016.

[83] S. Minucci and P. G. Pelicci. Histone deacetylase inhibitors and the promise of epigenetic (and more) treatments for cancer. *Nature Reviews Cancer*, 6 (1):38–51, 2006.

[84] E. A. Heller, P. D. Hsu, E. A. Heller, E. S. Lander, H. M. Cates, H. M. Cates, F. Zhang, C. J. Peña, C. J. Peña, H. Sun, H. Sun, N. Shao, N. Shao, J. Feng, J. Feng, S. A. Golden, S. A. Golden, J. P. Herman, J. P. Herman, J. J. Walsh, J. J. Walsh, M. Mazei-Robison, M. Mazei-Robison, D. Ferguson, D. Ferguson, S. Knight, S. Knight, M. A. Gerber, M. A. Gerber, C. Nievera, C. Nievera, M.-H. Han, M.-H. Han, S. J. Russo, S. J. Russo, C. S. Tamminga, C. S. Tamminga, R. L. Neve, R. L. Neve, L. Shen, L. Shen, H. S. Zhang, H. S. Zhang, F. Zhang, F. Zhang, E. J. Nestler, and E. J. Nestler. Locus-specific epigenetic remodeling controls addiction- and depression-related behaviors. *Nature Neuroscience*, 17(12):1720–1727, 2014.

[85] L. Bintu, J. Yong, Y. E. Antebi, K. McCue, Y. Kazuki, N. Uno, M. Oshimura, and M. B. Elowitz. Dynamics of epigenetic regulation at the single-cell level. *Science*, 351(6274):720–724, 2016.

[86] A. Amabile, A. Migliara, P. Capasso, M. Biffi, D. Cittaro, L. Naldini, and A. Lombardo. Inheritable Silencing of Endogenous Genes by Hit- and-Run Targeted Epigenetic Editing. *Cell*, 167(1):219–224.e14, 2016.

[87] X. S. Liu, H. Wu, X. Ji, Y. Stelzer, X. Wu, S. Czauderna, J. Shu, D. Dadon, R. A. Young, and R. Jaenisch. Editing DNA Methylation in the Mammalian Genome. *Cell*, 167(1):233–235.e17, 2016.

[88] D. L. Bernstein, J. E. Le Lay, E. G. Ruano, and K. H. Kaestner. TALE-mediated epigenetic suppression of CDKN2A increases replication in human fibroblasts. *Journal of Clinical Investigation*, 125(5):1998–2006, 2015.

[89] A. W. Snowden, P. D. Gregory, C. C. Case, and C. O. Pabo. Gene-specific targeting of H3K9 methylation is sufficient for initiating repression in vivo. *Current Biology*, 12(24):2159–2166, 2002.

[90] S. Konermann, M. D. Brigham, A. E. Trevino, P. D. Hsu, M. Heidenreich, L. Cong, R. J. Platt, D. A. Scott, G. M. Church, and F. Zhang. Optical control of mammalian endogenous transcription and epigenetic states. *Nature*, 500(7463):472–476, 2014.

[91] S. R. Choudhury, Y. Cui, K. Lubecka, B. Stefanska, and J. Irudayaraj. CRISPR-dCas9 mediated TET1 targeting for selective DNA demethylation at BRCA1 promoter. *Oncotarget*, 7(29):46545–46556, 2016.

[92] L. A. Gilbert, M. A. Horlbeck, B. Adamson, J. E. Villalta, Y. Chen, E. H. Whitehead, C. Guimaraes, B. Panning, H. L. Ploegh, M. C. Bassik, L. S. Qi, M. Kampmann, and J. S. Weissman. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*, 159:647–661, 2014.

[93] H. Koike-Yusa, Y. Li, E.-P. Tan, M. Del Castillo Velasco-Herrera, and K. Yusa. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature Biotechnology*, 32(3):267–273, 2014.

[94] O. Shalem, N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, T. Mikkelson, D. Heckl, B. L. Ebert, D. E. Root, J. G. Doench, and F. Zhang. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science*, 343: 84–87, 2013.

[95] T. Wang, J. J. Wei, D. M. Sabatini, and E. S. Lander. Genetic Screens in Human Cells Using the CRISPR/Cas9 System. *Science*, 343:80–84, 2013.

[96] Y. Zhou, S. Zhu, C. Cai, P. Yuan, C. Li, Y. Huang, and W. Wei. High-

throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature*, 509(7501):487–491, 2015.

[97] T. Hart, M. Chandrashekhar, M. Aregger, Z. Steinhart, K. R. Brown, G. MacLeod, M. Mis, M. Zimmermann, A. Fradet-Turcotte, S. Sun, P. Mero, P. Dirks, S. Sidhu, F. P. Roth, O. S. Rissland, D. Durocher, S. Angers, and J. Moffat. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell*, 163(6):1–13, 2015.

[98] S. M. Sidik, D. Huet, S. M. Ganesan, M.-H. Huynh, T. Wang, A. S. Nasamu, P. Thiru, J. P. J. Saeij, V. B. Carruthers, J. C. Niles, and S. Lourido. A Genome-wide CRISPR Screen in Toxoplasma Identifies Essential Apicomplexan Genes. *Cell*, 166(6):1423–1430.e12, 2016.

[99] M. E. Tanenbaum, L. A. Gilbert, L. S. Qi, J. S. Weissman, and R. D. Vale. A Protein-Tagging System for Signal Amplification in Gene Expression and Fluorescence Imaging. *Cell*, 159(3):635–646, 2014.

[100] M. C. Canver, E. C. Smith, F. Sher, L. Pinello, N. E. Sanjana, O. Shalem, D. D. Chen, P. G. Schupp, D. S. Vinjamur, S. P. Garcia, S. Luc, R. Kurita, Y. Nakamura, Y. Fujiwara, T. Maeda, G.-C. Yuan, F. Zhang, S. H. Orkin, and D. E. Bauer. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature*, 527(7577):192–197, 2015.

[101] G. Korkmaz, R. Lopes, A. P. Ugalde, E. Nevedomskaya, R. Han, K. Myacheva, W. Zwart, R. Elkon, and R. Agami. Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nature Biotechnology*, 34(2):192–198, 2016.

[102] S. Lamouille, J. Xu, and R. Derynck. Molecular mechanisms of epithelial–mesenchymal transition. *Nature Reviews Molecular Cell Biology*, 15(3): 178–196, 2014.

[103] J. P. Thiery, H. Acloque, R. Y. J. Huang, and M. A. Nieto. Epithelial-Mesenchymal Transitions in Development and Disease. *Cell*, 139(5):871–890, 2009.

[104] O. H. Ocaña, R. Córcoles, Á. Fabra, G. Moreno-Bueno, H. Acloque, S. Vega, A. Barrallo-Gimeno, A. Cano, and M. A. Nieto. Metastatic Colonization Requires the Repression of the Epithelial-Mesenchymal Transition Inducer Prrx1. *Cancer Cell*, 22(6):709–724, 2012.

[105] J. H. Tsai, J. L. Donaher, D. A. Murphy, S. Chau, and J. Yang. Spa-

tiotemporal Regulation of Epithelial-Mesenchymal Transition Is Essential for Squamous Cell Carcinoma Metastasis. *Cancer Cell*, 22(6):725–736, 2012.

[106] C. L. Chaffer and R. A. Weinberg. A Perspective on Cancer Cell Metastasis. *Science*, 331(6024):1559–1564, 2011.

[107] R. B. Hazan, G. R. Phillips, R. F. Qiao, L. Norton, and S. A. Aaronson. Exogenous expression of N-cadherin in breast cancer cells induces cell migration, invasion, and metastasis. *The Journal of Cell Biology*, 148(4): 779–790, 2000.

[108] A. K. Perl, P. Wilgenbus, U. Dahl, H. Semb, and G. Christofori. A causal role for E-cadherin in the transition from adenoma to carcinoma. *Nature*, 392(6672):190–193, 1998.

[109] G. Berx and F. van Roy. Involvement of Members of the Cadherin Superfamily in Cancer. *Cold Spring Harbor Perspectives in Biology*, 1(6): a003129–a003129, 2009.

[110] K. Vleminckx, L. Vakaet, M. Mareel, W. Fiers, and F. van Roy. Genetic manipulation of E-cadherin expression by epithelial tumor cells reveals an invasion suppressor role. *Cell*, 66(1):107–119, 1991.

[111] J. Yang and R. A. Weinberg. Epithelial-Mesenchymal Transition: At the Crossroads of Development and Tumor Metastasis. *Developmental Cell*, 14 (6):818–829, 2008.

[112] J. H. Taube, J. I. Herschkowitz, K. Komurov, A. Y. Zhou, S. Gupta, J. Yang, K. Hartwell, T. T. Onder, P. B. Gupta, K. W. Evans, B. G. Hollier, P. T. Ram, E. S. Lander, J. M. Rosen, R. A. Weinberg, and S. A. Mani. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *PNAS*, 107(35):15449–15454, 2010.

[113] H. Pages. *BSgenome: Infrastructure for Biostrings-based genome data packages*. R package version 1.28.0.

[114] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[115] Y. Namiki, T. Ishida, and Y. Akiyama. Acceleration of sequence clustering using longest common subsequence filtering. *BMC Bioinformatics*, 14 (Suppl 8):S7, 2013.

[116] H. Pages, P. Aboyoun, R. Gentleman, and S. DebRoy. *Biostrings: String objects representing biological sequences, and matching algorithms*. R package version 2.30.0.

[117] J. Sambrook and D. W. Russell. Isolation of DNA fragments from polyacrylamide gels by the crush and soak method. *CSH Protoc*, 2006(1), 2006.

[118] J. Sambrook and D. Russell. Purification of nucleic acids by extraction with phenol:chloroform. *CSH Protocols*, 1, 2006.

[119] D. G. Gibson, L. Young, R.-Y. Chuang, J. C. Venter, C. A. Hutchison, and H. O. Smith. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*, 6(5):343–345, 2009.

[120] M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*, 2011.

[121] `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`.

[122] A. B. Lane, M. Strzelecka, A. Ettinger, A. W. Grenfell, T. Wittmann, and R. Heald. Enzymatically Generated CRISPR Libraries for Genome Labeling and Screening. *Developmental Cell*, 34(3):373–378, 2015.

[123] B. De Craene and G. Berx. Regulatory networks defining EMT during cancer initiation and progression. *Nature Reviews Cancer*, 13(2):97–110, 2013.

[124] Y.-C. Chou. Variations in genome-wide RNAi screens: lessons from influenza research. *Journal of Clinical Bioinformatics*, pages 1–9, 2015.

[125] H. E. Davis, M. Rosinski, J. R. Morgan, and M. L. Yarmush. Charged polymers modulate retrovirus transduction via membrane charge neutralization and virus aggregation. *Biophysical Journal*, 86(2):1234–1242, 2004.

[126] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[127] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12): 550, 2014.

[128] `https://CRAN.R-project.org/package=desiR/`.

[129] M. Tahiliani, K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, and A. Rao. Conversion of 5-

Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science*, 324(5929):930–935, 2009.

[130] B. Xiao, B. Xiao, C. Jing, C. Jing, J. R. Wilson, J. R. Wilson, P. A. Walker, P. A. Walker, N. Vasisht, N. Vasisht, G. Kelly, G. Kelly, S. Howell, S. Howell, I. A. Taylor, I. A. Taylor, G. M. Blackburn, G. M. Blackburn, S. J. Gamblin, and S. J. Gamblin. Structure and catalytic mechanism of the human histone methyltransferase SET7/9. *Nature*, 421(6923):652–656, 2003.

[131] M. Yang, C. B. Gocke, X. Luo, D. Borek, D. R. Tomchick, M. Machius, Z. Otwinowski, and H. Yu. Structural Basis for CoREST-Dependent Demethylation of Nucleosomes by the Human LSD1 Histone Demethylase. *Molecular Cell*, 23(3):377–387, 2006.

[132] R. E. Collins, M. Tachibana, H. Tamaru, K. M. Smith, D. Jia, X. Zhang, E. U. Selker, Y. Shinkai, and X. Cheng. In Vitro and in Vivo Analyses of a Phe/Tyr Switch Controlling Product Specificity of Histone Lysine Methyltransferases. *Journal of Biological Chemistry*, 280(7):5563–5570, 2005.

[133] Z. Chen, J. Zang, J. Whetstine, X. Hong, F. Davrazou, T. G. Kutateladze, M. Simpson, Q. Mao, C.-H. Pan, S. Dai, J. Hagman, K. Hansen, Y. Shi, and G. Zhang. Structural Insights into Histone Demethylation by JMJD2 Family Members. *Cell*, 125(4):691–702, 2006.

[134] L. Balakrishnan, J. Stewart, P. Polaczek, J. L. Campbell, and R. A. Bambara. Acetylation of Dna2 Endonuclease/Helicase and Flap Endonuclease 1 by p300 Promotes DNA Stability by Creating Long Flap Intermediates. *Journal of Biological Chemistry*, 285(7):4398–4404, 2010.

[135] I. Gregoretti, Y.-M. Lee, and H. V. Goodson. Molecular Evolution of the Histone Deacetylase Family: Functional Implications of Phylogenetic Analysis. *Journal of Molecular Biology*, 338(1):17–31, 2004.

[136] A. L. Szymczak-Workman, K. M. Vignali, and D. A. A. Vignali. Design and Construction of 2A Peptide-Linked Multicistronic Vectors. *Cold Spring Harbor Protocols*, 2012(2):pdb.ip067876–pdb.ip067876, 2012.

[137] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

[138] F. A. Ran, P. D. Hsu, C.-Y. Lin, J. S. Gootenberg, S. Konermann, A. E.

Trevino, D. A. Scott, A. Inoue, S. Matoba, Y. Zhang, and F. Zhang. Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell*, 154(6):1380–1389, 2013.

[139] P. Cuatrecasas, S. Fuchs, and C. B. Anfinsen. Catalytic properties and specificity of the extracellular nuclease of Staphylococcus aureus. *The Journal of Biological Chemistry*, 242(7):1541–1547, 1967.

[140] S. Q. Tsai, Z. Zheng, N. T. Nguyen, M. Liebers, V. V. Topkar, V. Thapar, N. Wyvekens, C. Khayter, A. J. Iafrate, L. P. Le, M. J. Aryee, and J. K. Joung. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology*, 33(2):187–197, 2014.

[141] Y. Wang, J. Liu, X. Ying, P. C. Lin, and B. P. Zhou. Twist-mediated Epithelial-mesenchymal Transition Promotes Breast Tumor Cell Invasion viaInhibition of Hippo Pathway. *Scientific Reports*, 6:1–10, 2016.

[142] S. Guaita, I. Puig, C. Franci, M. Garrido, D. Dominguez, E. Batlle, E. Sancho, S. Dedhar, A. G. de Herreros, and J. Baulida. Snail Induction of Epithelial to Mesenchymal Transition in Tumor Cells Is Accompanied by MUC1 Repression and ZEB1 Expression. *Journal of Biological Chemistry*, 277(42):39209–39216, 2002.

[143] F. Fan, S. Samuel, K. W. Evans, J. Lu, L. Xia, Y. Zhou, E. Sceusi, F. Tozzi, X.-C. Ye, S. A. Mani, and L. M. Ellis. Overexpression of Snail induces epithelial-mesenchymal transition and a cancer stem cell-like phenotype in human colorectal cancer cells. *Cancer Medicine*, 1(1):5–16, 2012.

[144] V. Bolos. The transcription factor Slug represses E-cadherin expression and induces epithelial to mesenchymal transitions: a comparison with Snail and E47 repressors. *Journal of Cell Science*, 116(3):499–511, 2002.

[145] K. Vuoriluoto, H. Haugen, S. Kiviluoto, J.-P. Mpindi, J. Nevo, C. Gjerdrum, C. Tiron, J. B. Lorens, and J. Ivaska. Vimentin regulates EMT induction by Slug and oncogenic H-Ras and migration by governing Axl expression in breast cancer. *Oncogene*, 30(12):1436–1448, 2010.

[146] W. L. Tam, W. L. Tam, R. A. Weinberg, and R. A. Weinberg. The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nature Medicine*, 19 (11):1438–1449, 2013.

[147] S. J. Serrano-Gomez, M. Maziveyi, and S. K. Alahari. Regulation of

epithelial-mesenchymal transition through epigenetic and post- translational modifications. *Molecular Cancer*, 15(18):1–14, 2016.

[148] K. Best, T. Oakes, J. M. Heather, J. Shawe-Taylor, and B. Chain. Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Scientific Reports*, 5:1–13, 2015.

[149] T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1):72–74, 2011.

[150] J. A. Casbon, R. J. Osborne, S. Brenner, and C. P. Lichtenstein. A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Research*, 39(12):e81–e81, 2011.

[151] N. Thomas, J. Heather, W. Ndifon, J. Shawe-Taylor, and B. Chain. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics*, 29(5):542–550, 2013.

[152] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[153] W. Li, H. Xu, T. Xiao, L. Cong, M. I. Love, F. Zhang, R. A. Irizarry, J. S. Liu, M. Brown, and X. S. Liu. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biology*, 15(554):1–12, 2015.

[154] A. Derti, P. Garrett-Engele, K. D. MacIsaac, R. C. Stevens, S. Sriram, R. Chen, C. A. Rohl, J. M. Johnson, and T. Babak. A quantitative atlas of polyadenylation in five mammals. *Genome Research*, 22(6):1173–1183, 2012.

[155] F. Bushman, M. Lewinski, A. Ciuffi, S. Barr, J. Leipzig, S. Hannenhalli, and C. Hoffmann. Genome-wide analysis of retroviral DNA integration. *Nature Reviews Microbiology*, 3(11):848–858, 2005.

[156] M. Kumar, B. Keller, N. Makalou, and R. E. Sutton. Systematic determination of the packaging limit of lentiviral vectors. *Human Gene Therapy*, 12(15):1893–1905, 2001.

[157] A. Stockinger, A. Eger, J. Wolf, H. Beug, and R. Foisner. E-cadherin regulates cell growth by modulating proliferation-dependent β-catenin transcriptional activity. *The Journal of Cell Biology*, 154(6):1185–1196, 2001.

[158] J. Liu, G. Hu, D. Chen, A.-Y. Gong, G. S. Soori, T. J. Dobleman, and X.-M. Chen. Suppression of SCARA5 by Snail1 is essential for EMT-associated cell migration of A549 cells. *Oncogenesis*, 2(9):e73–10, 2013.

[159] `http://www.biostars.org/p/14614/`.

[160] P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6):276–277, 2000.

[161] `http://hannonlab.cshl.edu/fastx_toolkit/`.

[162] C. Berry, S. Hannenhalli, J. Leipzig, and F. D. Bushman. Selection of Target Sites for Mobile DNA Integration in the Human Genome. *PLoS Comput Biol*, 2(11):e157, 2006.

[163] `http://www.biostars.org/p/1195/`.

[164] `http://www.biostars.org/p/6219/`.

[165] P. Mali, J. Aach, P. B. Stranges, K. M. Esvelt, M. Moosburner, S. Kosuri, L. Yang, and G. M. Church. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature Biotechnology*, 31(9):833–838, 2013.

[166] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

Appendices

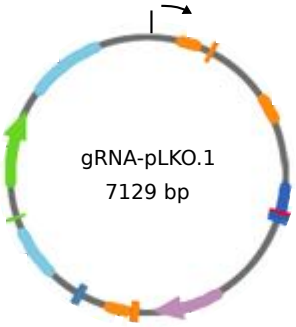
Appendix A

Plasmid maps and Supplementary tables

A.1 gRNA expression plasmids

Plasmid gRNA-pLKO.1:

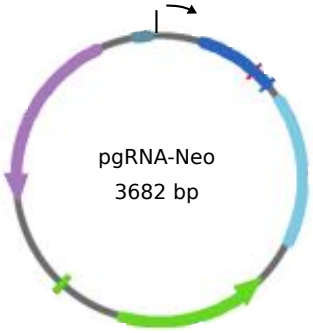
235 - 415: HIV-1 5 LTR
526 - 570: HIV-1 psi pack
1080 - 1313: RRE
1840 - 2089: U6 promoter
2089 - 2089: gRNA - transcr. start
2090 - 2165: scaffold
2166 - 2172: terminator
2928 - 3527: PuroR
3637 - 3659: cPPT
3655 - 3707: delta U3
3708 - 3888: HIV-1 3 LTR
4121 - 4198: SV40 ORI
4468 - 4908: F1 ori
4979 - 5007: Amp prom
5247 - 5906: AmpR
6058 - 6686: ColE1 origin



Sequence:

AGCTTAATGTAGTCTTATGCAATACTCTTGTAGTCTTGCAACATGGTAACGATGAGTTAGCAACATGCCTTACAAGGAGAGAAAAAGCACCGTGCATGCCGATTGGTGGAAGTAAGGTGGTACGATCGT
GCCTTATTAGGAAGGCAACAGACGGGTCTGACATGGATTGGACGAACCACTGAATTGCCGCATTGCAGAGATATTGTATTTAAGTGCCTAGCTCGATACATAAACGGGTCTCTCTGGTTAGACCAGATC
TGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCCACTGCTTAAGCCTCAATAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGTTGTGTGACTCTGGTAACTAGAGATCCCTCAGACCC
TTTTAGTCAGTGTGGAAAATCTCTAGCAGTGGCGCCCGAACAGGGACTTGAAAGCGAAAGGGAAACCAGAGGAGCTCTCTCGACGCAGGACTCGGCTTGCTGAAGCGCGCACGGCAAGAGGCGAGGGGC
GGCGACTGGTGAGTACGCCAAAAATTTTGACTAGCGGAGGCTAGAAGGAGAGAGATGGGTGCGAGAGCGTCAGTATTAAGCGGGGGAGAATTAGATCGCGATGGGAAAAAATTCGGTTAAGGCCAGGGG
GAAAGAAAAAATATAAATTAAAACATATAGTATGGGCAAGCAGGGAGCTAGAACGATTCGCAGTTAATCCTGGCCTGTTAGAAACATCAGAAGGCTGTAGACAAATACTGGGACAGCTACAACCATCCC
TTCAGACAGGATCAGAAGAACTTAGATCATTATATAATACAGTAGCAACCCTCTATTGTGTGCATCAAAGGATAGAGATAAAAGACACCAAGGAAGCTTTAGACAAGATAGAGGAAGAGCAAAACAAAA
GTAAGACCACCGCACAGCAAGCGGCCGCTGATCTTCAGACCTGGAGGAGGAGATATGAGGGACAATTGGAGAAGTGAATTATATAAATATAAAGTAGTAAAAATTGAACCATTAGGAGTAGCACCCACC
AAGGCAAAGAGAAGAGTGGTGCAGAGAGAAAAAAGAGCAGTGGGAATAGGAGCTTTGTTCCTTGGGTTCTTGGGAGCAGCAGGAAGCACTATGGGCGCAGCGTCAATGACGCTGACGGTACAGGCCAGA
CAATTATTGTCTGGTATAGTGCAGCAGCAGAACAATTTGCTGAGGGCTATTGAGGCGCAACAGCATCTGTTGCAACTCACAGTCTGGGGCATCAAGCAGCTCCAGGCAAGAATCCTGGCTGTGGAAAGA
TACCTAAAGGATCAACAGCTCCTGGGGATTTGGGGTTGCTCTGGAAAACTCATTTGCACCACTGCTGTGCCTTGGAATGCTAGTTGGAGTAATAAATCTCTGGAACAGATTTGGAATCACACGACCTGG
ATGGAGTGGGACAGAGAAATTAACAATTACACAAGCTTAATACACTCCTTAATTGAAGAATCGCAAAACCAGCAAGAAAAGAATGAACAAGAATTATTGGAATTAGATAAATGGGCAAGTTTGTGGAAT
TGGTTTAACATAACAAATTGGCTGTGGTATATAAAATTATTCATAATGATAGTAGGAGGCTTGGTAGGTTTAAGAATAGTTTTTGCTGTACTTTCTATAGTGAATAGAGTTAGGCAGGGATATTCACCA
TTATCGTTTCAGACCCACCTCCCAACCCCGAGGGGACCCGACAGGCCCGAAGGAATAGAAGAAGAAGGTGGAGAGAGAGACAGAGACAGATCCATTCGATTAGTGAACGGATCTCGACGGTATCGATCA
CGAGACTAGCCTCGAGCGGCCGCCCCCTTCACCGAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTAGAGAGATAATTGGAATTAATTTGACTGTAAACACAAAGATA
TTAGTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATAT
ATCTTGTGGAAAGGACGAAACACCGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTTAAGCTTGGGCCGCTCGAGGTA
CTCGACCTCGAGACAAATGGCAGTATTCATCCACAATTTTAAAAGAAAAGGGGGGATTGGGGGGTACAGTGCAGGGGAAAGAATAGTAGACATAATAGCAACAGACATACAAACTAAAGAATTACAAAA
ACAAATTACAAAAATTCAAAATTTTCGGGTTTATTACAGGGACAGCAGAGATCCACTTTGGCCGCGGCTCGAGGGGGTTGGGGTTGCGCCTTTTCCAAGGCAGCCCTGGGTTTGCGCAGGGACGCGGCT
GCTCTGGGCGTGGTTCCGGGAAACGCAGCGGCGCCGACCCTGGGTCTCGCACATTCTTCACGTCCGTTCGCAGCGTCACCCGGATCTTCGCCGCTACCCTTGTGGGCCCCCCGGCGACGCTTCCTGCTC
CGCCCCTAAGTCGGGAAGGTTCCTTGCGGTTCGCGGCGTGCCGGACGTGACAAACGGAAGCCGCACGTCTCACTAGTACCCTCGCAGACGGACAGCGCCAGGGAGCAATGGCAGCGCGCCGACCGCGAT
GGGCTGTGGCCAATAGCGGCTGCTCAGCAGGGCGCGCCGAGAGCAGCGGCCGGGAAGGGGCGGTGCGGGAGGCGGGGTGTGGGGCGGTAGTGTGGGCCCTGTTCCTGCCCGCGCGGTGTTCCGCATTCT
GCAAGCCTCCGGAGCGCACGTCGGCAGTCGGCTCCCTCGTTGACCGAATCACCGACCTCTCTCCCCAGGGGGATCCACCGGAGCTTACCATGACCGAGTACAAGCCCACGGTGCGCCTCGCCACCCGCG
ACGACGTCCCCAGGGCCGTACGCACCCTCGCCGCCGCGTTCGCCGACTACCCCGCCACGCGCCACACCGTCGATCCGGACCGCCACATCGAGCGGGTCACCGAGCTGCAAGAACTCTTCCTCACGCGCG
TCGGGCTCGACATCGGCAAGGTGTGGGTCGCGGACGACGGCGCCGCGGTGGCGGTCTGGACCACGCCGGAGAGCGTCGAAGCGGGGGCGGTGTTCGCCGAGATCGGCCCGCGCATGGCCGAGTTGAGCG
GTTCCCGGCTGGCCGCGCAGCAACAGATGGAAGGCCTCCTGGCGCCGCACCGGCCCAAGGAGCCCGCGTGGTTCCTGGCCACCGTCGGCGTCTCGCCCGACCACCAGGGCAAGGGTCTGGGCAGCGCCG
TCGTGCTCCCCGGAGTGGAGGCGGCCGAGCGGCGCCGGGGTGCCCGCCTTCCTGGAGACCTCCGCGCCCCGCAACCTCCCCTTCTACGAGCGGCTCGGCTTCACCGTCACCGCCGACGTCGAGGTGCCCG
AAGGACCGCGCACCTGGTGCATGACCCGCAAGCCCGGTGCCTGACGCCCGCCCCACGACCCGCAGCGCCCGACCGAAAGGAGCGCACGACCCCATGCATCGGTACCTTTAAGACCAATGACTTACAAGG
CAGCTGTAGATCTTAGCCACTTTTTAAAAGAAAAGGGGGGACTGGAAGGGCTAATTCACTCCCAACGAAGACAAGATCTGCTTTTTGCTTGTACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGG
AGCTCTCTGGCTAACTAGGGAACCCACTGCTTAAGCCTCAATAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGTTGTGTGACTCTGGTAACTAGAGATCCCTCAGACCCTTTTAGTCAG
TGTGGAAAATCTCTAGCAGTAGTAGTTCATGTCATCTTATTATTCAGTATTTATAACTTGCAAAGAAATGAATATCAGAGAGTGAGAGGAACTTGTTTATTGCAGCTTATAATGGTTACAAATAAAGCA
ATAGCATCACAAATTTCACAAATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATCTTATCATGTCTGGCTCTAGCTATCCCGCCCCTAACTCCGCCCATCCCGCC
CCTAACTCCGCCCAGTTCCGCCCATTCTCCGCCCCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCGGCCTCTGAGCTATTCCAGAAGTAGTGAGGAGGCTTTTTTGGAGGCCT
AGGGACGTACCCAATTCGCCCTATAGTGAGTCGTATTACGCGCGCTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTC
GCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAATGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGC
GTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGT
GCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTC
CAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATA
TTAACGCTTACAATTTAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAAT
ATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGT
TGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCC
GTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTG
CCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGC
TGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGG
AGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGC
CCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCATATA
TACTTTAGATTGATTTAAAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAA
AGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAA
CTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTG
CCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAAC
TGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCT
GGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTT
GCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCG
AGGAAGCGGAAGAGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTT
AGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGCTATGACCATGATTACGCCAAGCGCGCAATTA
ACCCTCACTAAAGGGAACAAAAGCTGGAGCTGCA

Plasmid pgRNA-Neo:
160 - 422: U6 promoter
423 - 423: gRNA - transcr. start
424 - 500: scaffold
501 - 506: terminator
585 - 1213: ColE1 origin
1365 - 2024: AmpR
2264 - 2292: Amp prom
2656 - 3447: Kan/neoR
3589 - 3666: SV40 ORI

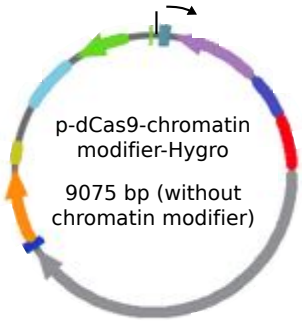


Sequence:

GGGACTATGGTTGCTGACTAATTGAGATGCATGCTTTGCATACTTCTGCCTGCTGGGGAGCCTGGGGACTTTCCACACCTGGTTGCTGACTAATTGAGATGCACTGTACAAAAAAGCAGGCTTTAAAGG
AACCAATTCAGTCGACTGGATCCGGTACCAAGGTCGGGCAGGAAGAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAA
ACACAAAGATATTAGTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCT
TGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTTAAGCTTGGGC
CGCTCGAGGTACCTCTCTACATATGACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGA
CGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCT
TCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAAC
TATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGC
TACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAG
CAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACC
TAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCC
ATAGTTGCCTGACTCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCA
GCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCT
ACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATC
GTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGA
GAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATC
TTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATA
AGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTT
CCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTCGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATT
TGTAACCATTATAAGCTGCAATAAACAAGTTGGGGTGGGCGAAGAACTCCAGCATGAGATCCCCGCGCTGGAGGATCATCCAGCCGGCGTCCCGGAAAACGATTCCGAAGCCCAACCTTTCATAGAAGG
CGGCGGTGGAATCGAAATCTCGTGATGGCAGGTTGGGCGTCGCTTGGTCGGTCATTTCGAACCCCAGAGTCCCGCTCAGAAGAACTCGTCAAGAAGGCGATAGAAGGCGATGCGCTGCGAATCGGGAGC
GGCGATACCGTAAAGCACGAGGAAGCGGTCAGCCCATTCGCCGCCAAGCTCTTCAGCAATATCACGGGTAGCCAACGCTATGTCCTGATAGCGGTCCGCCACACCCAGCCGGCCACAGTCGATGAATCC
AGAAAAGCGGCCATTTTCCACCATGATATTCGGCAAGCAGGCATCGCCATGGGTCACGACGAGATCCTCGCCGTCGGGCATGCGCGCCTTGAGCCTGGCGAACAGTTCGGCTGGCGCGAGCCCCTGATG
CTCTTCGTCCAGATCATCCTGATCGACAAGACCGGCTTCCATCCGAGTACGTGCTCGCTCGATGCGATGTTTCGCTTGGTGGTCGAATGGGCAGGTAGCCGGATCAAGCGTATGCAGCCGCCGCATTGC
ATCAGCCATGATGGATACTTTCTCGGCAGGAGCAAGGTGAGATGACAGGAGATCCTGCCCCGGCACTTCGCCCAATAGCAGCCAGTCCCTTCCCGCTTCAGTGACAACGTCGAGCACAGCTGCGCAAGG
AACGCCCGTCGTGGCCAGCCACGATAGCCGCGCTGCCTCGTCCTGCAGTTCATTCAGGGCACCGGACAGGTCGGTCTTGACAAAAAGAACCGGGCGCCCCTGCGCTGACAGCCGGAACACGGCGGCATC
AGAGCAGCCGATTGTCTGTTGTGCCCAGTCATAGCCGAATAGCCTCTCCACCCAAGCGGCCGGAGAACCTGCGTGCAATCCATCTTGTTCAATCATGCGAAACGATCCTCATCCTGTCTCTTGATCAGA
TCCGAAAATGGATATACAAGCTCCCGGGAGCTTTTTGCAAAAGCCTAGGCCTCCAAAAAAGCCTCCTCACTACTTCTGGAATAGCTCAGAGGCAGAGGCGGCCTCGGCCTCTGCATAAATAAAAAAAAT
TAGTCAGCCATGGGGCGGAGAATGGGCGGAACTGGGCGGAGTTAGGGGCGGGATGGGCGGAGTTAGGGGC

A.2 dCas9-chromatin modifier expression plasmids

Plasmid p-dCas9-chromatin-modifier-Hygro:
77 - 207: CMV polyA
238 - 1263: HygromycinR
1281 - 1782: PGK promoter
1786 - 2388: CMV promoter
2442 - 6578: dCas9
6579 - 6645: 3X FLAG
6651 - 6665: G3S linker
6666 - 6666: chromatin modifier
6688 - 6915: bGH pA
7370 - 7998: ColE origin
8150 - 8809: AmpR
9049 - 9077: Amp promoter

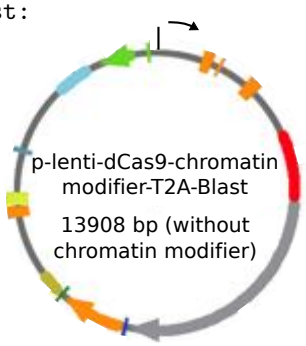


Sequence:

TGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTCGACGGGTATACAGACATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAG
TGAAAAAAATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTGGGGTGGGCGAAGAACTCTCGGCATCTACTCTATTCCTTTGCCCTCGGACG
AGTGCTGGGGCGTCGGTTTCCACTATCGGCGAGTACTTCTACACAGCCATCGGTCCAGACGGCCGCGCTTCTGCGGGCGATTTGTGTACGCCCGACAGTCCCGGCTCCGGATCGGACGATTGCGTCGCA
TCGACCCTGCGCCCAAGCTGCATCATCGAAATTGCCGTCAACCAAGCTCTGATAGAGTTGGTCAAGACCAATGCGGAGCATATACGCCCGGAGCCGCGGCGATCCTGCAAGCTCCGGATGCCTCCGCTC
GAAGTAGCGCGTCTGCTGCTCCATACAAGCCAACCACGGCCTCCAGAAGAAGATGTTGGCGACCTCGTATTGGGAATCCCCGAACATCGCCTCGCTCCAGTCAATGACCGCTGTTATGCGGCCATTGTC
CGTCAGGACATTGTTGGAGCCGAAATCCGCGTGCACGAGGTGCCGGACTTCGGGGCAGTCCTCGGCCCAAAGCATCAGCTCATCGAGAGCCTGCGCGACGGACGCACTGACGGTGTCGTCCATCACAGT
TTGCCAGTGATACACATGGGGATCAGCAATCGCGCATATGAAATCACGCCATGTAGTGTATTGACCGATTCCTTGCGGTCCGAATGGGCCGAACCCGCTCGTCTGGCTAAGATCGGCCGCAGCGATCGC
ATCCATGGCCTCCGCGACCGGCTGCAGAACAGCGGGCAGTTCGGTTTCAGGCAGGTCTTGCAACGTGACACCCTGTGCACGGCGGGAGATGCAATAGGTCAGGCTCTCGCTGAATTCCCCAATGTCAAG
CACTTCCGGAATCGGGAGCGCGGCCGATGCAAAGTGCCGATAAACATAACGATCTTTGTAGAAACCATCGGCGCAGCTATTTACCCGCAGGACATATCCACGCCCTCCTACATCGAAGCTGAAAGCACG
AGATTCTTCGCCCTCCGAGAGCTGCATCAGGTCGGAGACGCTGTCGAACTTTTCGATCAGAAACTTCTCGACAGACGTCGCGGTGAGTTCAGGCTTTTTCATGGTGGCGGGATGCAGGTCGAAAGGCCC
GGAGATGAGGAAGAGGAGAACAGCGCGGCAGACGTGCGCTTTTGAAGCGTGCAGAATGCCGGGCCTCCGGAGGACCTTCGGGCGCCCGCCCCGCCCCTGAGCCCGCCCCCGGACCCA
CCCCTTCCCAGCCTCTGAGCCCAGAAAGCGAAGGAGCAAAGCTGCTATTGGCCGCTGCCCCAAAGGCCTACCCGCTTCCATTGCTCAGCGGTGCTGTCCATCTGCACGAGACTAGTGAGACGTGCTACT
TCCATTTGTCACGTCCTGCACGACGCGAGCTGCGGGGCGGGGGGGAACTTCCTGACTAGGGGAGGAGTAGAAGGTGGCGCGAAGGGGCCACCAAAGAACGGAGCCGGTTGGCGCCTACCGGTGGATGTG
GAATGTGTGCGAGGCCAGAGGCCACTTGTGTAGCGCCAAGTGCCCAGCGGGGCTGCTAAAGCGCATGCTCCAGACTGCCTTGGGAAAAGCGCCTCCCCTACCCGGTTGACATTGATTATTGACTAGTTA
TTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGAC
GTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAA
TGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGG
GCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCA
AATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTGGTTTAGTGAACCGTCAGATCCGCTAGAGATCCGCGGCCGCTAATACGACTCACTATAGGGAGAGCCGCCACCATGGATAAAA
AGTATTCTATTGGTTTAGCCATCGGCACTAATTCCGTTGGATGGGCTGTCATAACCGATGAATACAAAGTACCTTCAAAGAAATTTAAGGTGTTGGGGAACACAGACCGTCATTCGATTAAAAGAATC
TTATCGGTGCCCTCCTATTCGATAGTGGCGAAACGGCAGAGGCGACTCGCCTGAAACGAACCGCTCGGAGAAGGTATACACGTCGCAAGAACCGAATATGTTACTTACAAGAAATTTTTAGCAATGAGA
TGGCCAAAGTTGACGATTCTTTCTTTCACCGTTTGGAAGAGTCCTTCCTTGTCGAAGAGGACAAGAAACATGAACGGCACCCCATCTTTGGAAACATAGTAGATGAGGTGGCATATCATGAAAAGTACC
CAACGATTTATCACCTCAGAAAAAAGCTAGTTGACTCAACTGATAAAGCGGACCTGAGGTTAATCTACTTGGCTCTTGCCCATATGATAAAGTTCCGTGGGCACTTTCTCATTGAGGGTGATCTAAATC
CGGACAACTCGGATGTCGACAAACTGTTCATCCAGTTAGTACAAACCTATAATCAGTTGTTTGAAGAGAACCCTATAAATGCAAGTGGCGTGGATGCGAAGGCTATTCTTAGCGCCCGCCTCTCTAAAT
CCCGACGGCTAGAAAACCTGATCGCACAATTACCCGGAGAGAAGAAAAATGGGTTGTTCGGTAACCTTATAGCGCTCTCACTAGGCCTGACACCAAATTTTAAGTCGAACTTCGACTTAGCTGAAGATG
CCAAATTGCAGCTTAGTAAGGACACGTACGATGACGATCTCGACAATCTACTGGCACAAATTGGAGATCAGTATGCGGACTTATTTTTGGCTGCCAAAAACCTTAGCGATGCAATCCTCCTATCTGACA
TACTGAGAGTTAATACTGAGATTACCAAGGCGCCGTTATCCGCTTCAATGATCAAAAGGTACGATGAACATCACCAAGACTTGACACTTCTCAAGGCCCTAGTCCGTCAGCAACTGCCTGAGAAATATA
AGGAAATATTCTTTGATCAGTCGAAAAACGGGTACGCAGGTTATATTGACGGCGGAGCGAGTCAAGAGGAATTCTACAAGTTTATCAAACCCATATTAGAGAAGATGGATGGGACGGAAGAGTTGCTTG
TAAAACTCAATCGCGAAGATCTACTGCGAAAGCAGCGGACTTTCGACAACGGTAGCATTCCACATCAAATCCACTTAGGCGAATTGCATGCTATACTTAGAAGGCAGGAGGATTTTTATCCGTTCCTCA
AAGACAATCGTGAAAAGATTGAGAAAATCCTAACCTTTCGCATACCTTACTATGTGGGACCCCTGGCCCGAGGGAACTCTCGGTTCGCATGGATGACAAGAAAGTCCGAAGAAACGATTACTCCATGGA
ATTTTGAGGAAGTTGTCGATAAAGGTGCGTCAGCTCAATCGTTCATCGAGAGGATGACCAACTTTGACAAGAATTTACCGAACGAAAAAGTATTGCCTAAGCACAGTTTACTTTACGAGTATTTCACAG
TGTACAATGAACTCACGAAAGTTAAGTATGTCACTGAGGGCATGCGTAAACCCGCCTTCTAAGCGGAGAACAGAAGAAAGCAATAGTAGATCTGTTATTCAAGACCAACCGCAAAGTGACAGTTAAGC
AATTGAAAGAGGACTACTTTAAGAAAATTGAATGCTTCGATTCTGTCGAGATCTCCGGGGTAGAAGATCGATTTAATGCGTCACTTGGTACGTATCATGACCTCCTAAAGATAATTAAAGATAAGGACT
TCCTGGATAACGAAGAGAATGAAGATATCTTAGAAGATATAGTGTTGACTCTTACCCTCTTTGAAGATCGGGAAATGATTGAGGAAAGACTAAAAACATACGCTCACCTGTTCGACGATAAGGTTATGA
AACAGTTAAAGAGGCGTCGCTATACGGGCTGGGGACGATTGTCGCGGAAACTTATCAACGGGATAAGAGACAAGCAAAGTGGTAAAACTATTCTCGATTTTCTAAAGAGCGACGGCTTCGCCAATAGGA
ACTTTATGCAGCTGATCCATGATGACTCTTTAACCTTCAAAGAGGATATACAAAAGGCACAGGTTTCCGGACAAGGGGACTCATTGCACGAACATATTGCGAATCTTGCTGGTTCGCCAGCCATCAAAA
AGGGCATACTCCAGACAGTCAAAGTAGTGGATGAGCTAGTTAAGGTCATGGGACGTCACAAACCGGAAAACATTGTAATCGAGATGGCACGCGAAAATCAAACGACTCAGAAGGGGCAAAAAAACAGTC
GAGAGCGGATGAAGAGAATAGAAGAGGGTATTAAAGAACTGGGCAGCCAGATCTTAAAGGAGCATCCTGTGGAAAATACCCAATTGCAGAACGAGAAACTTTACCTCTATTACCTACAAAATGGAAGGG
ACATGTATGTTGATCAGGAACTGGACATAAACCGTTTATCTGATTACGACGTCGATGCCATTGTACCCCAATCCTTTTTGAAGGACGATTCAATCGACAATAAAGTGCTTACACGCTCGGATAAGAACC
GAGGGAAAAGTGACAATGTTCCAAGCGAGGAAGTCGTAAAGAAAATGAAGAACTATTGGCGGCAGCTCCTAAATGCGAAACTGATAACGCAAAGAAAGTTCGATAACTTAACTAAAGCTGAGAGGGGTG
GCTTGTCTGAACTTGACAAGGCCGGATTTATTAAACGTCAGCTCGTGGAAACCCGCCAAATCACAAAGCATGTTGCACAGATACTAGATTCCCGAATGAATACGAAATACGACGAGAACGATAAGCTGA
TTCGGGAAGTCAAAGTAATCACTTTAAAGTCAAAATTGGTGTCGGACTTCAGAAAGGATTTTCAATTCTATAAAGTTAGGGAGATAAATAACTACCACCATGCGCACGACGCTTATCTTAATGCCGTCG
TAGGGACCGCACTCATTAAGAAATACCCGAAGCTAGAAAGTGAGTTTGTGTATGGTGATTACAAAGTTTATGACGTCCGTAAGATGATCGCGAAAAGCGAACAGGAGATAGGCAAGGCTACAGCCAAAT
ACTTCTTTTATTCTAACATTATGAATTTCTTTAAGACGGAAATCACTCTGGCAAACGGAGAGATACGCAAACGACCTTTAATTGAAACCAATGGGGAGACAGGTGAAATCGTATGGGATAAGGGCCGGG
ACTTCGCGACGGTGAGAAAAGTTTTGTCCATGCCCCAAGTCAACATAGTAAAGAAAACTGAGGTGCAGACCGGAGGGTTTTCAAAGGAATCGATTCTTCCAAAAAGGAATAGTGATAAGCTCATCGCTC
GTAAAAAGGACTGGGACCCGAAAAAGTACGGTGGCTTCGATAGCCCTACAGTTGCCTATTCTGTCCTAGTAGTGGCAAAAGTTGAGAAGGGAAAATCCAAGAAACTGAAGTCAGTCAAAGAATTATTGG
GGATAACGATTATGGAGCGCTCGTCTTTTGAAAAGAACCCCATCGACTTCCTTGAGGCGAAAGGTTACAAGGAAGTAAAAAAGGATCTCATAATTAAACTACCAAAGTATAGTCTGTTTGAGTTAGAAA
ATGGCCGAAAACGGATGTTGGCTAGCGCCGGAGAGCTTCAAAAGGGGAACGAACTCGCACTACCGTCTAAATACGTGAATTTCCTGTATTTAGCGTCCCATTACGAGAAGTTGAAAGGTTCACCTGAAG
ATAACGAACAGAAGCAACTTTTGTTGAGCAGCACAAACATTATCTCGACGAAATCATAGAGCAAATTTCGGAATTCAGTAAGAGAGTCATCCTAGCTGATGCCAATCTGGACAAAGTATTAAGCGCAT
ACAACAAGCACAGGGATAAACCCATACGTGAGCAGGCGGAAAATATTATCCATTTGTTTACTCTTACCAACCTCGGCGCTCCAGCCGCATTCAAGTATTTTGACACAACGATAGATCGCAAACGATACA
CTTCTACCAAGGAGGTGCTAGACGCGACACTGATTCACCAATCCATCACGGGATTATATGAAACTCGGATAGATTTGTCACAGCTTGGGGGTGACGGATCCCCCAAGAAGAAGAGGAAAGTCTCGAGCG
ACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAGGCTGCAGGAGGCGGAGGTAGC---insert-chromatin-modifier-sequence---TGA
AAACCCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTG
CATCGCATTGTCTGAGTAGGTGTCATTCTATTCTGGGGGGTGGGGTGGGGCAGGACAGCAAGGGGGAGGATTGGGAAGACAATAGCAGGCATGCTGGGGATGCGGTGGGCTCTATGGCTTCTGAGGCGG
AAAGAACCAGCTGGGGCTCGATACCGTCGACCTCTAGCTAGAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAA
AGTGTAAAGCCTAGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAG
GCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATA
ACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGA
GGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGG
CGCTTTCTCAATGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGT
CCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGA
CAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCA
GAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAA
ATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGAC
TCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCG
AGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGG
TGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTA
AGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGC
GGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGA
GATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGA
AATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAA

Plasmid lenti-dCas9-chromatin-modifier-T2A-Blast:

835 - 1015: HIV-1 5 LTR
1126 - 1170: HIV-1 psi pack
1680 - 1913: RRE
2694 - 3865: EF1a promoter
3878 - 8014: dCas9
8057 - 8080: 3X FLAG
8102 - 8102: Chromatin modifier
8115 - 8177: T2A
8178 - 8576: BlasticidinR
9714 - 9894: HIV-1 3 LTR
9923 - 10150: bGH PA
10818 - 10895: SV40 ORI
12139 - 12767: ColE1 origin
12919 - 13578: AmpR
13818 - 13846: Amp prom



Sequence:

GTCGACGGATCGGGAGATCTCCCGATCCCCTATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATCTGCTCCCTGCTTGTGTGTTGGAGGTCGCTGAGTAGTGCGCGAGCA
AAATTTAAGCTACAACAAGGCAAGGCTTGACCGACAATTGCATGAAGAATCTGCTTAGGGTTAGGCGTTTTGCGCTGCTTCGCGATGTACGGGCCAGATATACGCGTTGACATTGATTATTGACTAGTT
ATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGA
CGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCA
ATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATG
GGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGC
AAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGCGCGTTTTGCCTGTACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCCACTGCTTAAG
CCTCAATAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGTTGTGTGACTCTGGTAACTAGAGATCCCTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTAGCAGTGGCGCCCGAACAGGG
ACTTGAAAGCGAAAGGGAAACCAGAGGAGCTCTCTCGACGCAGGACTCGGCTTGCTGAAGCGCGCACGGCAAGAGGCGAGGGGCGGCGACTGGTGAGTACGCCAAAAATTTTGACTAGCGGAGGCTAGA
AGGAGAGAGATGGGTGCGAGAGCGTCAGTATTAAGCGGGGGAGAATTAGATCGCGATGGGAAAAAATTCGGTTAAGGCCAGGGGGAAAGAAAAAATATAAATTAAAACATATAGTATGGGCAAGCAGGG
AGCTAGAACGATTCGCAGTTAATCCTGGCCTGTTAGAAACATCAGAAGGCTGTAGACAAATACTGGGACAGCTACAACCATCCCTTCAGACAGGATCAGAAGAACTTAGATCATTATATAATACAGTAG
CAACCCTCTATTGTGTGCATCAAAGGATAGAGATAAAAGACACCAAGGAAGCTTTAGACAAGATAGAGGAAGAGCAAAACAAAAGTAAGACCACCGCACAGCAAGCGGCCGCTGATCTTCAGACCTGGA
GGAGGAGATATGAGGGACAATTGGAGAAGTGAATTATATAAATATAAAGTAGTAAAAATTGAACCATTAGGAGTAGCACCCACCAAGGCAAAGAGAAGAGTGGTGCAGAGAGAAAAAAGAGCAGTGGGA
ATAGGAGCTTTGTTCCTTGGGTTCTTGGGAGCAGCAGGAAGCACTATGGGCGCAGCGTCAATGACGCTGACGGTACAGGCCAGACAATTATTGTCTGGTATAGTGCAGCAGCAGAACAATTTGCTGAGG
GCTATTGAGGCGCAACAGCATCTGTTGCAACTCACAGTCTGGGGCATCAAGCAGCTCCAGGCAAGAATCCTGGCTGTGGAAAGATACCTAAAGGATCAACAGCTCCTGGGGATTTGGGGTTGCTCTGGA
AAACTCATTTGCACCACTGCTGTGCCTTGGAATGCTAGTTGGAGTAATAAATCTCTGGAACAGATTTGGAATCACACGACCTGGATGGAGTGGGACAGAGAAATTAACAATTACACAAGCTTAATACAC
TCCTTAATTGAAGAATCGCAAAACCAGCAAGAAAAGAATGAACAAGAATTATTGGAATTAGATAAATGGGCAAGTTTGTGGAATTGGTTTAACATAACAAATTGGCTGTGGTATATAAAATTATTCATA
ATGATAGTAGGAGGCTTGGTAGGTTTAAGAATAGTTTTTGCTGTACTTTCTATAGTGAATAGAGTTAGGCAGGGATATTCACCATTATCGTTTCAGACCCACCTCCCAACCCCGAGGGGACCCGACAGG
CCCGAAGGAATAGAAGAAGAAGGTGGAGAGAGAGACAGAGACAGATCCATTCGATTAGTGAACGGATCGGCACTGCGTGCGCCAATTCTGCAGACAAATGGCAGTATTCATCCACAATTTTAAAAGAAA
AGGGGGGATTGGGGGGTACAGTGCAGGGGAAAGAATAGTAGACATAATAGCAACAGACATACAAACTAAAGAATTACAAAAACAAATTACAAAAATTCAAAATTTTCGGGTTTATTACAGGGACAGCAG
AGATCCAGTTTGGTTAATTAGCTAGCTGCAAAGATGGATAAAGTTTTAAACAGAGAGGAATCTTTGCAGCTAATGGACCTTCTAGGTCTTGAAAGGAGTGGGAATTGGCTCCGGTGCCCGTCAGTGGGC
AGAGCGCACATCGCCCACAGTCCCCGAGAAGTTGGGGGGAGGGGTCGGCAATTGAACCGGTGCCTAGAGAAGGTGGCGCGGGGTAAACTGGGAAAGTGATGTCGTGTACTGGCTCCGCCTTTTTCCCGA
GGGTGGGGGAGAACCGTATATAAGTGCAGTAGTCGCCGTGAACGTTCTTTTTCGCAACGGGTTTGCCGCCAGAACACAGGTAAGTGCCGTGTGTGGTTCCCGCGGGCCTGGCCTCTTTACGGGTTATGG
CCCTTGCGTGCCTTGAATTACTTCCACCTGGCTGCAGTACGTGATTCTTGATCCCGAGCTTCGGGTTGGAAGTGGGTGGGAGAGTTCGAGGCCTTGCGCTTAAGGAGCCCCTTCGCCTCGTGCTTGAGT
TGAGGCCTGGCCTGGGCGCTGGGGCCGCCGCGTGCGAATCTGGTGGCACCTTCGCGCCTGTCTCGCTGCTTTCGATAAGTCTCTAGCCATTTAAAATTTTTGATGACCTGCTGCGACGCTTTTTTTCTG
GCAAGATAGTCTTGTAAATGCGGGCCAAGATCTGCACACTGGTATTTCGGTTTTTGGGGCCGCGGGCGGCGACGGGGCCCGTGCGTCCCAGCGCACATGTTCGGCGAGGCGGGGCCTGCGAGCGCGGCC
ACCGAGAATCGGACGGGGGTAGTCTCAAGCTGGCCGGCCTGCTCTGGTGCCTGGCCTCGCGCCGCCGTGTATCGCCCCGCCCTGGGCGGCAAGGCTGGCCCGGTCGGCACCAGTTGCGTGAGCGGAAAG
ATGGCCGCTTCCCGGCCCTGCTGCAGGGAGCTCAAAATGGAGGACGCGGCGCTCGGGAGAGCGGGCGGGTGAGTCACCCACACAAAGGAAAAGGGCCTTTCCGTCCTCAGCCGTCGCTTCATGTGACTC
CACGGAGTACCGGGCGCCGTCCAGGCACCTCGATTAGTTCTCGAGCTTTTGGAGTACGTCGTCTTTAGGTTGGGGGGAGGGGTTTTATGCGATGGAGTTTCCCCACACTGAGTGGGTGGAGACTGAAGT
TAGGCCAGCTTGGCACTTGATGTAATTCTCCTTGGAATTTGCCCTTTTTGAGTTTGGATCTTGGTTCATTCTCAAGCCTCAGACAGTGGTTCAAAGTTTTTTTCTTCCATTTCAGGTGTCGTGACGTAC
GGCCACCATGGATAAAAAGTATTCTATTGGTTTAGCCATCGGCACTAATTCCGTTGGATGGGCTGTCATAACCGATGAATACAAAGTACCTTCAAAGAAATTTAAGGTGTTGGGGAACACAGACCGTCA
TTCGATTAAAAAGAATCTTATCGGTGCCCTCCTATTCGATAGTGGCGAAACGGCAGAGGCGACTCGCCTGAAACGAACCGCTCGGAGAAGGTATACACGTCGCAAGAACCGAATATGTTACTTACAAGA
AATTTTTAGCAATGAGATGGCCAAAGTTGACGATTCTTTCTTTCACCGTTTGGAAGAGTCCTTCCTTGTCGAAGAGGACAAGAAACATGAACGGCACCCCATCTTTGGAAACATAGTAGATGAGGTGGC
ATATCATGAAAAGTACCCAACGATTTATCACCTCAGAAAAAAGCTAGTTGACTCAACTGATAAAGCGGACCTGAGGTTAATCTACTTGGCTCTTGCCCATATGATAAAGTTCCGTGGGCACTTTCTCAT
TGAGGGTGATCTAAATCCGGACAACTCGGATGTCGACAAACTGTTCATCCAGTTAGTACAAACCTATAATCAGTTGTTTGAAGAGAACCCTATAAATGCAAGTGGCGTGGATGCGAAGGCTATTCTTAG
CGCCCGCCTCTCTAAATCCCGACGGCTAGAAAACCTGATCGCACAATTACCCGGAGAGAAGAAAAATGGGTTGTTCGGTAACCTTATAGCGCTCTCACTAGGCCTGACACCAAATTTTAAGTCGAACTT
CGACTTAGCTGAAGATGCCAAATTGCAGCTTAGTAAGGACACGTACGATGACGATCTCGACAATCTACTGGCACAAATTGGAGATCAGTATGCGGACTTATTTTTGGCTGCCAAAAACCTTAGCGATGC
AATCCTCCTATCTGACATACTGAGAGTTAATACTGAGATTACCAAGGCGCCGTTATCCGCTTCAATGATCAAAAGGTACGATGAACATCACCAAGACTTGACACTTCTCAAGGCCCTAGTCCGTCAGCA
ACTGCCTGAGAAATATAAGGAAATATTCTTTGATCAGTCGAAAAACGGGTACGCAGGTTATATTGACGGCGGAGCGAGTCAAGAGGAATTCTACAAGTTTATCAAACCCATATTAGAGAAGATGGATGG
GACGGAAGAGTTGCTTGTAAAACTCAATCGCGAAGATCTACTGCGAAAGCAGCGGACTTTCGACAACGGTAGCATTCCACATCAAATCCACTTAGGCGAATTGCATGCTATACTTAGAAGGCAGGAGGA
TTTTTATCCGTTCCTCAAAGACAATCGTGAAAAGATTGAGAAAATCCTAACCTTTCGCATACCTTACTATGTGGGACCCCTGGCCCGAGGGAACTCTCGGTTCGCATGGATGACAAGAAAGTCCGAAGA
AACGATTACTCCATGGAATTTTGAGGAAGTTGTCGATAAAGGTGCGTCAGCTCAATCGTTCATCGAGAGGATGACCAACTTTGACAAGAATTTACCGAACGAAAAAGTATTGCCTAAGCACAGTTTACT
TTACGAGTATTTCACAGTGTACAATGAACTCACGAAAGTTAAGTATGTCACTGAGGGCATGCGTAAACCCGCCTTTCTAAGCGGAGAACAGAAGAAAGCAATAGTAGATCTGTTATTCAAGACCAACCG
CAAAGTGACAGTTAAGCAATTGAAAGAGGACTACTTTAAGAAAATTGAATGCTTCGATTCTGTCGAGATCTCCGGGGTAGAAGATCGATTTAATGCGTCACTTGGTACGTATCATGACCTCCTAAAGAT
AATTAAAGATAAGGACTTCCTGGATAACGAAGAGAATGAAGATATCTTAGAAGATATAGTGTTGACTCTTACCCTCTTTGAAGATCGGGAAATGATTGAGGAAAGACTAAAAACATACGCTCACCTGTT
CGACGATAAGGTTATGAAACAGTTAAAGAGGCGTCGCTATACGGGCTGGGGACGATTGTCGCGGAAACTTATCAACGGGATAAGAGACAAGCAAAGTGGTAAAACTATTCTCGATTTTCTAAAGAGCGA
CGGCTTCGCCAATAGGAACTTTATGCAGCTGATCCATGATGACTCTTTAACCTTCAAAGAGGATATACAAAAGGCACAGGTTTCCGGACAAGGGGACTCATTGCACGAACATATTGCGAATCTTGCTGG
TTCGCCAGCCATCAAAAAGGGCATACTCCAGACAGTCAAAGTAGTGGATGAGCTAGTTAAGGTCATGGGACGTCACAAACCGGAAAACATTGTAATCGAGATGGCACGCGAAAATCAAACGACTCAGAA
GGGGCAAAAAAACAGTCGAGAGCGGATGAAGAGAATAGAAGAGGGTATTAAAGAACTGGGCAGCCAGATCTTAAAGGAGCATCCTGTGGAAAATACCCAATTGCAGAACGAGAAACTTTACCTCTATTA
CCTACAAAATGGAAGGGACATGTATGTTGATCAGGAACTGGACATAAACCGTTTATCTGATTACGACGTCGATGCCATTGTACCCCAATCCTTTTTGAAGGACGATTCAATCGACAATAAAGTGCTTAC
ACGCTCGGATAAGAACCGAGGGAAAAGTGACAATGTTCCAAGCGAGGAAGTCGTAAAGAAAATGAAGAACTATTGGCGGCAGCTCCTAAATGCGAAACTGATAACGCAAAGAAAGTTCGATAACTTAAC
TAAAGCTGAGAGGGGTGGCTTGTCTGAACTTGACAAGGCCGGATTTATTAAACGTCAGCTCGTGGAAACCCGCCAAATCACAAAGCATGTTGCACAGATACTAGATTCCCGAATGAATACGAAATACGA
CGAGAACGATAAGCTGATTCGGGAAGTCAAAGTAATCACTTTAAAGTCAAAATTGGTGTCGGACTTCAGAAAGGATTTTCAATTCTATAAAGTTAGGGAGATAAATAACTACCACCATGCGCACGACGC
TTATCTTAATGCCGTCGTAGGGACCGCACTCATTAAGAAATACCCGAAGCTAGAAAGTGAGTTTGTGTATGGTGATTACAAAGTTTATGACGTCCGTAAGATGATCGCGAAAAGCGAACAGGAGATAGG
CAAGGCTACAGCCAAATACTTCTTTTATTCTAACATTATGAATTTCTTTAAGACGGAAATCACTCTGGCAAACGGAGAGATACGCAAACGACCTTTAATTGAAACCAATGGGGAGACAGGTGAAATCGT
ATGGGATAAGGGCCGGGACTTCGCGACGGTGAGAAAAGTTTTGTCCATGCCCCAAGTCAACATAGTAAAGAAAACTGAGGTGCAGACCGGAGGGTTTTCAAAGGAATCGATTCTTCCAAAAAGGAATAG
TGATAAGCTCATCGCTCGTAAAAAGGACTGGGACCCGAAAAAGTACGGTGGCTTCGATAGCCCTACAGTTGCCTATTCTGTCCTAGTAGTGGCAAAAGTTGAGAAGGGAAAATCCAAGAAACTGAAGTC
AGTCAAAGAATTATTGGGGATAACGATTATGGAGCGCTCGTCTTTTGAAAAGAACCCCATCGACTTCCTTGAGGCGAAAGGTTACAAGGAAGTAAAAAAGGATCTCATAATTAAACTACCAAAGTATAG

Sequence (continued):

TCTGTTTGAGTTAGAAAATGGCCGAAAACGGATGTTGGCTAGCGCCGGAGAGCTTCAAAAGGGGAACGAACTCGCACTACCGTCTAAATACGTGAATTTCCTGTATTTAGCGTCCCATTACGAGAAGTT
GAAAGGTTCACCTGAAGATAACGAACAGAAGCAACTTTTTGTTGAGCAGCACAAACATTATCTCGACGAAATCATAGAGCAAATTTCGGAATTCAGTAAGAGAGTCATCCTAGCTGATGCCAATCTGGA
CAAAGTATTAAGCGCATACAACAAGCACAGGGATAAACCCATACGTGAGCAGGCGGAAAATATTATCCATTTGTTTACTCTTACCAACCTCGGCGCTCCAGCCGCATTCAAGTATTTTGACACAACGAT
AGATCGCAAACGATACACTTCTACCAAGGAGGTGCTAGACGCGACACTGATTCACCAATCCATCACGGGATTATATGAAACTCGGATAGATTTGTCACAGCTTGGGGGTGACGGATCCCCCAAGAAGAA
GAGGAAAGTCTCGAGCGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAGGCTGCAGGAGGCGGAGGTAGC---insert-chromatin-modifi
er-sequence---ATTAACTGTACAGGCAGTGGAGAGGGCAGAGGAAGTCTGCTAACATGCGGTGACGTCGAGGAGAATCCTGGCCCAATGGCCAAGCCTTTGTCTCAAGAAGAATCCACCCTCATTG
AAAGAGCAACGGCTACAATCAACAGCATCCCCATCTCTGAAGACTACAGCGTCGCCAGCGCAGCTCTCTCTAGCGACGGCCGCATCTTCACTGGTGTCAATGTATATCATTTTACTGGGGGACCTTGTG
CAGAACTCGTGGTGCTGGGCACTGCTGCTGCTGCGGCAGCTGGCAACCTGACTTGTATCGTCGCGATCGGAAATGAGAACAGGGGCATCTTGAGCCCCTGCGGACGGTGCCGACAGGTGCTTCTCGATC
TGCATCCTGGGATCAAAGCCATAGTGAAGGACAGTGATGGACAGCCGACGGCAGTTGGGATTCGTGAATTGCTGCCCTCTGGTTATGTGTGGGAGGGCTAAGAATTCGATATCAAGCTTATCGGTAATC
AACCTCTGGATTACAAAATTTGTGAAAGATTGACTGGTATTCTTAACTATGTTGCTCCTTTTACGCTATGTGGATACGCTGCTTTAATGCCTTTGTATCATGCTATTGCTTCCCGTATGGCTTTCATTT
TCTCCTCCTTGTATAAATCCTGGTTGCTGTCTCTTTATGAGGAGTTGTGGCCCGTTGTCAGGCAACGTGGCGTGGTGTGCACTGTGTTTGCTGACGCAACCCCCACTGGTTGGGGCATTCGCCACCACCT
GTCAGCTCCTTTCCGGGACTTTCGCTTTCCCCCTCCCTATTGCCACGGCGGAACTCATCGCCGCCTGCCTTGCCCGCTGCTGGACAGGGGCTCGGCTGTTGGGCACTGACAATTCCGTGGTGTTGTCGG
GGAAATCATCGTCCTTTCCTTGGCTGCTCGCCTGTGTTGCCACCTGGATTCTGCGCGGGACGTCCTTCTGCTACGTCCCTTCGGCCCTCAATCCAGCGGACCTTCCTTCCCGCGGCCTGCTGCCGGCTC
TGCGGCCTCTTCCGCGTCTTCGCCTTCGCCCTCAGACGAGTCGGATCTCCCTTTGGGCCGCCTCCCCGCGCATCGATACCGTCGACCTCGAGACCTAGAAAAACATGGAGCAATCACAAGTAGCAATACAG
CAGCTACCAATGCTGATTGTGCCTGGCTAGAAGCACAAGAGGAGGAGGAGGTGGGTTTTCCAGTCACACCTCAGGTACCTTTAAGACCAATGACTTACAAGGCAGCTGTAGATCTTAGCCACTTTTTAA
AAGAAAAGGGGGGACTGGAAGGGCTAATTCACTCCCAACGAAGACAAGATATCCTTGATCTGTGGATCTACCACACACAAGGCTACTTCCCTGATTGGCAGAACTACACACCAGGGCCAGGGATCAGAT
ATCCACTGACCTTTGGATGGTGCTACAAGCTAGTACCAGTTGAGCAAGAGAAGGTAGAAGAAGCCAATGAAGGAGAACACCCGCTTGTTACACCCTGTGAGCCTGCATGGGATGGATGACCCGAGAA
GAGAAGTATTAGAGTGGAGGTTTGACAGCCGCCTAGCATTTCATCACATGGCCCGAGAGCTGCATCCGGACTGTACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAG
GGAACCCACTGCTTAAGCCTCAATAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGTTGTGTGACTCTGGTAACTAGAGATCCCTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTAGC
AGGGCCCGTTTAAACCCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAA
TGAGGAAATTGCATCGCATTGTCTGAGTAGGTGTCATTCTATTCTGGGGGGTGGGGCAGGACAGCAAGGGGGAGGATTGGGAAGACAATAGCAGGCATGCTGGGGATGCGGTGGGCTCTATGGC
TTCTGAGGCGGAAAGAACCAGCTGGGGCTCTAGGGGGTATCCCCACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCC
CGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGA
TTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGT
CTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTAATTCTGTGGAATGTGTGTCAGTTAGGGTGTGGAAAGTCCCCA
GGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCAGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCATA
GTCCCGCCCCTAACTCCGCCCATCCCGCCCCTAACTCCGCCCAGTTCCGCCCATTCTCCGCCCCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCTGCCTCTGAGCTATTCCAG
AAGTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGCTCCCGGGAGCTTGTATATCCATTTTCGGATCTGATCAGCACGTGTTGACAATTAATCATCGGCATAGTATATCGGCATAGTAT
AATACGACAAGGTGAGGAACTAAACCATGGCCAAGTTGACCAGTGCCGTTCCGGTGCTCACCGCGCGCGACGTCGCCGGAGCGGTCGAGTTCTGGACCGACCGGCTCGGGTTCTCCCGGGACTTCGTGG
AGGACGACTTCGCCGGTGTGGTCCGGGACGACGTGACCCTGTTCATCAGCGCGGTCCAGGACCAGGTGGTGCCGGACAACACCCTGGCCTGGGTGTGGGTGCGCGGCCTGGACGAGCTGTACGCCGAGT
GGTCGGAGGTCGTGTCCACGAACTTCCGGGACGCCTCCGGGCCGGCCATGACCGAGATCGGCGAGCAGCCGTGGGGGCGGGAGTTCGCCCTGCGCGACCCGGCCGGCAACTGCGTGCACTTCGTGGCCG
AGGAGCAGGACTGACACGTGCTACGAGATTTCGATTCCACCGCCGCCTTCTATGAAAGGTTGGGCTTCGGAATCGTTTTCCGGGACGCCGGCTGGATGATCCTCCAGCGCGGGGATCTCATGCTGGAGT
TCTTCGCCCACCCCAACTTGTTTATTGCAGCTTATAATGGTTACAAATAAAGCAATAGCATCACAAATTTCACAAATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAATG
TATCTTATCATGTCTGTATACCGTCGACCTCTAGCTAGAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAAAGT
GTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCG
GTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACG
CAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGT
GGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGC
TTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCA
ACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAG
TATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAA
AAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATT
AAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCC
CCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGC
GCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGT
CACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGT
TGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGC
GACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGAT
CCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAAT
GTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCC
GAAAAGTGCCACCTGAC

Chromatin modifier sequences:

DNMT3A (amino acids 598-908):

AACCACGACCAGGAATTTGACCCTCCAAAGGTTTACCCACCTGTCCCAGCTGAGAAGAGGAAGCCCATCCGGGTGCTGTCTCTCTTTGATGGAATCGGCTACAGGGCTCCTGGTGCTGAAGGACTTGGGC
ATTCAGGTGGACCGCTACATTGCCTCGGAGGTGTGTGAGGACTCCATCACGGTGGGCATGGTGCGGCACCAGGGGAAGATCATGTACGTCGGGGACGTCCGCAGCGTCACACAGAAGCATATCCAGGAG
TGGGGCCCATTCGATCTGGTGATTGGGGGCAGTCCCTGCAATGACCTCTCCATCGTCAACCCTGCTCGCAAGGGCCTCTACGAGGGCACTGGCCGGCTCTTCTTTGAGTTCTACCGCCTCCTGCATGAT
GCGCGGCCCAAGGAGGGAGATGATCGCCCCTTCTTCTGGCTCTTTGAGAATGTGGTGGCCATGGGCGTTAGTGACAAGAGGGACATCTCGCGATTTCTCGAGTCCAACCCTGTGATGATTGATGCCAAA
GAAGTGTCAGCTGCACACAGGGCCCGCTACTTCTGGGGTAACCTTCCCGGTATGAACAGGCCGTTGGCATCCACTGTGAATGATAAGCTGGAGCTGCAGGAGTGTCTGGAGCATGGCAGGATAGCCAAG
TTCAGCAAAGTGAGGACCATTACTACGAGGTCAAACTCCATAAAGCAGGGCAAAGACCAGCATTTTCCTGTCTTCATGAATGAGAAAGAGGACATCTTATGGTGCACTGAAATGGAAAGGGTATTTGGT
TTCCCAGTCCACTATACTGACGTCTCCAACATGAGCCGCTTGGCGAGGCAGAGACTGCTGGGCCGGTCATGGAGCGTGCCAGTCATCCGCCACCTCTTCGCTCCGCTGAAGGAGTAT

TET1 (amino acids 1418-2136):

GAACTGCCCACCTGCAGCTGTCTTGATCGAGTTATACAAAAAGACAAAGGCCCATATTATACACACCTTGGGGCAGGACCAAGTGTTGCTGCTGTCAGGGAAATCATGGAGAATAGGTATGGTCAAAAA
GGAAACGCAATAAGGATAGAAATAGTAGTGTACACCGGTAAAGAAGGGAAAAGCTCTCATGGGTGTCCAATTGCTAAGTGGGTTTTAAGAAGAAGCAGTGATGAAGAAAAAGTTCTTTGTTTGGTCCGG
CAGCGTACAGGCCACCACTGTCCAACTGCTGTGATGGTGGTGCTCATCATGGTGTGGGATGGCATCCCTCTTCCAATGGCCGACCGGCTATACACACAGAGCTCACAGAGAATCTAAAGTCATACAATGGG
CACCCTACCGACAGAAGATGCACCCTCAATGAAAATCGTACCTGTACATGTCAAGGAATTGATCCAGAGACTTGTGGAGCTTCATTCTCTTTTGGCTGTTCATGGAGTATGTACTTTAATGGCTGTAAG
TTTGGTAGAAGCCCAAGCCCCAGAAGATTTAGAATTGATCCAAGCTCTCCCTTACATGAAAAAAACCTTGAAGATAACTTACAGAGTTTGGCTACACGATTAGCTCCAATTTATAAGCAGTATGCTCCA
GTAGCTTACCAAAATCAGGTGGAATATGAAAATGTTGCCCGAGAATGTCGGCTTGGCAGCAAGGAAGGTCGTCCCTTCTCTGGGGTCACTGCTTGCCTGGACTTCTGTGCTCATCCCCACAGGGACATT
CACAACATGAATAATGGAAGCACTGTGGTTTGTACCTTAACTCGAGAAGATAACCGCTCTTTGGGTGTTATTCCTCAAGATGAGCAGCTCCATGTGCTACCTCTTTATAAGCTTTCAGACACAGATGAG
TTTGGCTCCAAGGAAGGAATGGAAGCCAAGATCAAATCTGGGGCCATCGAGGTCCTGGCACCCCGCCGCAAAAAAAGAACGTGTTTCACTCAGCCTGTTCCCCGTTCTGGAAAGAAGAGGGCTGCGATG
ATGACAGAGGTTCTTCGACATAAGATAAGGGCAGTGGAAAAGAAACCTATTCCCCGAATCAAGCGGAAGAATAACTCAACAACAACAAACAACAGTAAGCCTTCGTCACTGCCAACCTTAGGGAGTAAC
ACTGAGACCGTGCAACCTGAAGTAAAAAGTGAAACCGAACCCCATTTTATCTTAAAAAGTTCAGACAACACTAAAACTTATTCGCTGATGCCATCCGCTCCTCACCCAGTGAAAGAGGCATCTCCAGGC
TTCTCCTGGTCCCCGAAGACTGCTTCAGCCACACCAGCTCCACTGAAGAATGACGCAACAGCCTCATGCGGGTTTTCAGAAAGAAGCAGCACTCCCCACTGTACGATGCCTTCGGGAAGACTCAGTGGT
GCCAATGCAGCTGCTGCTGATGGCCCTGGCATTTCACAGCTTGGCGAAGTGGCTCCTCTCCCCACCCTGTCTGCTCCTGTGATGGAGCCCCTCATTAATTCTGAGCCTTCCACTGGTGTGACTGAGCCG
CTAACGCCTCATCAGCCAAACCACCAGCCCTCCTTCCTCACCTCTCCTCAAGACCTTGCCTCTTCTCCAATGGAAGAAGATGAGCAGCATTCTGAAGCAGATGAGCCTCCATCAGACGAACCCCTATCT
GATGACCCCCTGTCACCTGCTGAGGAGAAATTGCCCCACATTGATGAGTATTGGTCAGACAGTGAGCACATCTTTTTGGATGCAAATATTGGTGGGGTGGCCATCGCACCTGCTCACGGCTCGGTTTTG
ATTGAGTGTGCCCGGCGAGAGCTGCACGCTACCACTCCTGTTGAGCACCCCAACCGTAATCATCCAACCCGCCTCTCCCTTGTCTTTTACCAGCACAAAAACCTAAATAAGCCCCAACATGGTTTTGAA
CTAAACAAGATTAAGTTTGAGGCTAAAGAAGCTAAGAATAAGAAAATGAAGGCCTCAGAGCAAAAAGACCAGGCAGCTAATGAAGGTCCAGAACAGTCCTCTGAAGTAAATGAATTGAACCAAATTCCT
TCTCATAAAGCATTAACATTAACCCATGACAATGTTGTCACCGTGTCCCCTTATGCTCTCACACACGTTGCGGGGCCCTATAACCATTGGGTC

Chromatin modifier sequences (continued):

SET7 (amino acids 52-366):

TTCTTCTTTGATGGCAGCACCCTGGAGGGGTATTATGTGGATGATGCCTTGCAGGGCCAGGGAGTTTACACTTACGAAGATGGGGGAGTTCTCCAGGGCACGTATGTAGACGGAGAGCTGAACGGTCCA
GCCCAGGAATATGACACAGATGGGAGACTGATCTTCAAGGGGCAGTATAAAGATAACATTCGTCATGGAGTGTGCTGGATATATTACCCAGATGGAGGAAGCCTTGTAGGAGAAGTAAATGAAGATGGG
GAGATGACTGGAGAGAAGATAGCCTATGTGTACCCTGATGAGAGGACCGCACTTTATGGGAAATTTATTGATGGAGAGATGATAGAAGGCAAACTGGCTACCCTTATGTCCACTGAAGAAGGGAGGCCT
CACTTTGAACTGATGCCTGGAAATTCAGTGTACCACTTTGATAAGTCGACTTCATCTTGCATTTCTACCAATGCTCTTCTTCCAGATCCTTATGAATCAGAAAGGGTTTATGTTGCTGAATCTCTTATT
TCCAGTGCTGGAGAAGGACTTTTTTCAAAGGTAGCTGTGGGACCTAATACTGTTATGTCTTTTTATAATGGAGTTCGAATTACACACCAAGAGGTTGACAGCAGGGACTGGGCCCTTAATGGGAACACC
CTCTCCCTTGATGAAGAAACGGTCATTGATGTGCCTGAGCCCTATAACCACGTATCCAAGTACTGTGCCTCCTTGGGACACAAGGCAAATCACTCCTTCACTCCAAACTGCATCTACGATATGTTTGTC
CACCCCCGTTTTGGGCCCATCAAATGCATCCGCACCCTGAGAGCAGTGGAGGCCGATGAAGAGCTCACCGTTGCCTATGGCTATGACCACAGCCCCCCCGGGAAGAGTGGGCCTGAAGCCCCTGAGTGG
TACCAGGTGGAGCTGAAGGCCTTCCAGGCCACCCAGCAAAAG

LSD1 (amino acids 171-852):

CCATCGGGTGTGGAGGGCGCAGCTTTCCAGAGCCGACTTCCTCATGACCGGATGACTTCTCAAGAAGCAGCCTGTTTTCCAGATATTATCAGTGGACCACAACAGACCCAGAAGGTTTTTCTTTTCATT
AGAAACCGCACACTGCAGTTGTGGTTGGATAATCCAAAGATTCAGCTGACATTTGAGGCTACTCTCCAACAATTAGAAGCACCTTATAACAGTGATACTGTGCTTGTCCACCGAGTTCACAGTTATTTA
GAGCGTCATGGTCTTATCAACTTCGGCATCTATAAGAGGATAAAACCCCTACCAACTAAAAAGACAGGAAAGGTAATTATTATAGGCTCTGGGGTCTCAGGCTTGGCAGCAGCTCGACAGTTACAAAGT
TTTGGAATGGATGTCACACTTTTGGAAGCCAGGGATCGTGTGGGTGGACGAGTTGCCACATTTCGCAAAGGAAACTATGTAGCTGATCTTGGAGCCATGGTGGTAACAGGTCTTGGAGGGAATCCTATG
GCTGTGGTCAGCAAACAAGTAAATATGGAACTGGCCAAGATCAAGCAAAAATGCCCACTTTATGAAGCCAACGGACAAGCTGTTCCTAAAGAGAAAGATGAAATGGTAGAGCAAGAGTTTAACCGGTTG
CTAGAAGCTACATCTTACCTTAGTCATCAACTAGACTTCAATGTCCTCAATAATAAGCCTGTGTCCCTTGGCCAGGCATTGGAAGTTGTCATTCAGTTACAAGAGAAGCATGTCAAAGATGAGCAGATT
GAACATTGGAAGAAGATAGTGAAAACTCAGGAAGAATTGAAAGAACTTCTTAATAAGATGGTAAATTTGAAAGAGAAAATTAAAGAACTCCATCAGCAATACAAAGAAGCATCTGAAGTAAAGCCACCC
AGAGATATTACTGCCGAGTTCTTAGTGAAAAGCAAACACAGGGATCTGACCGCCCTATGCAAGGAATATGATGAATTAGCTGAAACACAAGGAAAGCTAGAAGAAAAACTTCAGGAGTTGGAAGCGAAT
CCCCCAAGTGATGTATATCTCTCATCAAGAGACAGACAAATACTTGATTGGCATTTTGCAAATCTTGAATTTGCTAATGCCACACCTCTCTCAACTCTCTCCCTTAAGCACTGGGATCAGGATGATGAC
TTTGAGTTCACTGGCAGCCACCTGACAGTAAGGAATGGCTACTCGTGTGTGCCTGTGGCTTTAGCAGAAGGCCTAGACATTAAACTGAATACAGCAGTGCGACAGGTTCGCTACACGGCTTCAGGATGT
GAAGTGATAGCTGTGAATACCCGCTCCACGAGTCAAACCTTTATTTATAAATGCGACGCAGTTCTCTGTACCCTTCCCCTGGGTGTGCTGAAGCAGCAGCCACCAGCCGTTCAGTTTGTGCCACCTCTC
CCTGAGTGGAAAACATCTGCAGTCCAAAGGATGGGATTTGGCAACCTTAACAAGGTGGTGTTGTGTTTTGATCGGGTGTTCTGGGATCCAAGTGTCAATTTGTTCGGGCATGTTGGCAGTACGACTGCC
AGCAGGGGTGAGCTCTTCCTCTTCTGGAACCTCTATAAAGCTCCAATACTGTTGGCACTAGTGGCAGGAGAAGCTGCTGGTATCATGGAAAACATAAGTGACGATGTGATTGTTGGCCGATGCCTGGCC
ATTCTCAAAGGGATTTTTGGTAGCAGTGCAGTACCTCAGCCCAAAGAAACTGTGGTGTCTCGTTGGCGTGCTGATCCCTGGGCTCGGGGCTCTTATTCCTATGTTGCTGCAGGATCATCTGGAAATGAC
TATGATTTAATGGCTCAGCCAATCACTCCTGGCCCCTCGATTCCAGGTGCCCCACAGCCGATTCCACGACTCTTCTTTGCGGGAGAACATACGATCCGTAACTACCCAGCCACAGTGCATGGTGCTCTG
CTGAGTGGGCTGCGAGAAGCGGGAAGAATTGCAGACCAGTTTTTGGGGGCCATGTATACGCTGCCTCGCCAGGCCACACCAGGTGTTCCTGCACAGCAGTCCCCAAGCATG

G9a (amino acids 621-1000):

GGCTATGAGAACGTGCCCATTCCCTGTGTCAACGGTGTGGATGGGGAGCCCTGCCCTGAGGATTACAAGTACATCTCAGAGAACTGCGAGACGTCCACCATGAACATCGATCGCAACATCACCCACCTG
CAGCACTGCACGTGTGTGGACGACTGCTCTAGCTCCAACTGCCTGTGCGGCCAGCTCAGCATCCGGTGCTGGTATGACAAGGATGGGCGATTGCTCCAGGAATTTAACAAGATTGAGCCTCCGCTGATT
TTCGAGTGTAACCAGGCGTGCTCATGCTGGAGAAACTGCAAGAACCGGGTCGTACAGAGTGGCATCAAGGTGCGGCTACAGCTCTACCGAACAGCCAAGATGGGCTGGGGGGTCCGCGCCCTGCAGACC
ATCCCACAGGGGACCTTCATCTGCGAGTATGTCGGGGAGCTGATCTCTGATGCTGAGGCTGATGTGAGAGAGGATGATTCTTACCTCTTCGACTTAGACAACAAGGATGGAGAGGTGTACTGCATAGAT
GCCCGTTACTATGGCAACATCAGCCGCTTCATCAACCACCTGTGTGACCCCAACATCATTCCCGTCCGGGTCTTCATGCTGCACCAAGACCTGCGATTTCCACGCATCGCCTTCTTCAGTTCCCGAGAC
ATCCGGACTGGGGAGGAGCTAGGGTTTGACTATGGCGACCGCTTCTGGGACATCAAAAGCAAATATTTCACCTGCCAATGTGGCTCTGAGAAGTGCAAGCACTCAGCCGAAGCCATTGCCCTGGAGCAG
AGCCGTCTGGCCCGCCTGGACCCACACCCTGAGCTGCTGCCCGAGCTCGGCTCCCTGCCCCCTGTCAACACA

JMJD2A (amino acids 1-350):

GCTTCTGAGTCTGAAACTCTGAATCCCAGTGCTAGGATAATGACCTTTTATCCAACTATGGAAGAGTTCCGAAACTTCAGTAGATACATTGCCTACATTGAATCCCAAGGAGCTCATCGGGCAGGGCTA
GCCAAGGTTGTTCCTCCAAAAGAGTGGAAGCCACGAGCATCCTATGATGACATTGATGATTTGGTCATTCCTGCCCCCATTCAACAGCTGGTGACGGGGCAGTCTGGCCTCTTTACTCAGTACAACATA
CAGAAGAAAGCCATGACTGTTCGAGAGTTCCGCAAGATAGCCAATAGCGATAAGTACTGTACCCCACGCTATAGTGAGTTTGAAGAGCTCGAGCGGAAATACTGGAAAAATCTTACATTCAATCCTCCA
ATCTATGGTGCAGATGTGAATGGTACCCTCTATGAAAAGCATGTTGATGAGTGGAATATTGGCCGGCTGAGAACAATCCTGGACTTGGTGGAAAAGGAGAGTGGGATCACCATTGAGGGTGTGAACACC
CCATACCTGTACTTTGGCATGTGGAAGACATCCTTTGCTTGGCACACTGAAGACATGGACCTCTACAGCATCAACTACCTGCACTTTGGAGAACCAAAGTCCTGGTACTCTGTTCCACCTGAGCATGGA
AAGCGGTTGGAACGCCTCGCCAAAGGCTTTTTCCCAGGAAGTGCTCAAAGCTGTGAGGCATTTCTCCGCCACAAGATGACCCTGATTTCCCCGTTAATGCTGAAGAAATATGGAATTCCCTTTGACAAG
GTGACTCAAGAGGCTGGAGAGTTTATGATCACTTTCCCTTATGGTTACCATGCCGGCTTTAACCATGGTTTTAACTGTGCGGAGTCTACCAATTTTGCTACCCGTCGGTGGATTGAGTACGGCAAGCAA
GCTGTGCTGTGCTCCTGTAGAAAGGACATGGTGAAGATCTCCATGGATGTGTTTGTGAGAAAGTTCCAGCCAGAAAGGTACAAACTTTGGAAAGCTGGGAAGGACAACACAGTTATTGACCATACTCTG
CCCACGCCAGAAGCA

p300 (amino acids 1284-1673):

AAAGAAAATAAGTTTTCTGCTAAAAGGTTGCCATCTACCAGACTTGGCACCTTTCTAGAGAATCGTGTGAATGACTTTCTGAGGCGACAGAATCACCCTGAGTCAGGAGAGGTCACTGTTAGAGTAGTT
CATGCTTCTGACAAAACCGTGGAAGTAAAACCAGGCATGAAAGCAAGGTTTGTGGACAGTGGAGAGATGGCAGAATCCTTTCCATACCGAACCAAAGCCCTCTTTGCCTTTGAAGAAATTGATGGTGTT
GACCTGTGCTTCTTTGGCATGCATGTTCAAGAGTATGGCTCTGACTGCCCTCCACCCAACCAGAGGAGAGTATACATATCTTACCTCGATAGTGTTCATTTCTTCCGTCCTAAATGCTTGAGGACTGCA
GTCTATCATGAAATCCTAATTGGATATTTAGAATATGTCAAGAAATTAGGTTACACAACAGGGCATATTTGGGCATGTCCACCAAGTGAGGGAGATGATTATATCTTCCATTGCCATCCTCCTGACCAG
AAGATACCCAAGCCCAAGCGACTGCAGGAATGGTACAAAAAAAATGCTTGACAAGGCTGTATCAGAGCGTATTGTCCATGACTACAAGGATATTTTTAAACAAGCTACTGAAGATAGATTAACAAGTGCA
AAGGAATTGCCTTATTTCGAGGGTGATTTCTGGCCCAATGTTCTGGAAGAAAGCATTAAGGAACTGGAACAGGAGGAAGAAGAGAGAAAACGAGAGGAAAACACCAGCAATGAAAGCACAGATGTGACC
AAGGGAGACAGCAAAAATGCTAAAAAGAAGAATAATAAGAAAACCAGCAAAAATAAGAGCAGCCTGAGTAGGGGCAACAAGAAGAAACCCGGGATGCCCAATGTATCTAACGACCTCTCACAGAAACTA
TATGCCACCATGGAGAAGCATAAAGAGGTCTTCTTTTGTGATCCGCCTCATTGCTGGCCCTGCTGCCAACTCCCTGCCTCCCATTGTTGATCCTGATCCTCTCATCCCCTGCGATCTGATGGATGGTCGG
GATGCGTTTCTCACGCTGGCAAGGGACAAGCACCTGGAGTTCTCTTCACTCCGAAGAGCCCAGTGGTCCACCATGTGCATGCTGGTGGAGCTGCACACGCAGAGCCAGGACCGCTTTGTCTACACCTGC
AATGAATGC

HDAC1 (amino acids 4-384):

ACGCAGGGCACCCGGAGGAAAGTCTGTTACTACTACGACGGGGATGTTGGAAATTACTATTATGGACAAGGCCACCCAATGAAGCCTCACCGAATCCGCATGACTCATAATTTGCTGCTCAACTATGGT
CTCTACCGAAAAATGGAAATCTATCGCCCTCACAAAGCCAATGCTGAGGAGATGACCAAGTACCACAGCGATGACTACATTAAATTCTTGCGCTCCATCCGTCCAGATAACATGTCGGAGTACAGCAAG
CAGATGCAGAGATTCAACGTTGGTGAGGACTGTCCAGTATTCGATGGCCTGTTTGAGTTCTGTCAGTTGTCTACTGGTGGTTCTGTGGCAAGTGCTGTGAAACTTAATAAGCAGCAGACGGACATCGCT
GTGAATTGGGCTGGGGGCCTGCACCATGCAAAGAAGTCCGAGGCATCTGGCTTCTGTTACGTCAATGATATCGTCTTGGCCATCCTGGAACTGCTAAAGTATCACCAGAGGGTGCTGTACATTGACATT
GATATTCACCATGGTGACGGCGTGGAAGAGGCCTTCTACACCACGGACCGGGTCATGACTGTGTCCTTTCATAAGTATGGAGAGTACTTCCCAGGAACTGGGGACCTACGGGATATCGGGGCTGGCAAA
GGCAAGTATTATGCTGTTAACTACCCGCTCCGAGACGGGATTGATGACGAGTCCTATGAGGCCATTTTCAAGCCGGTCATGTCCAAAGTAATGGAGATGTTCCAGCCTAGTGCGGTGGTCTTACAGTGT
GGCTCAGACTCCCTATCTGGGGATCGGTTAGGTTGCTTCAATCTAACTATCAAAGGACACGCCAAGTGTGTGGAATTTGTCAAGAGCTTTAACCTGCCTATGCTGATGCTGGGAGGCGGTGGTTACACC
ATTCGTAACGTTGCCCGGTGCTGGACATATGAGACAGCTGTGGCCCTGGATACGGAGATCCCTAATGAGCTTCCATACAATGACTACTTTGAATACTTTGGACCAGATTTCAAGCTCCACATCAGTCCT
TCCAATATGACTAACCAGAACACGAATGAGTACCTGGAGAAGATCAAACAGCGACTGTTTGAGAACCTTAGAATGCTGCCGCACGCACCTGGGGTCCAAATGCAGGCGATT

A.3 Data analysis: Read tables for sequencing of the degenerate and MNase digest libraries

Library name	Raw read count	Adapter trimmed	Reverse complement adapter trimmed	Reads after 4-way adapter trimming
255-N19-R1	938,334	319,270	281,424	600,694
256-C190-T20	1,143,513	449,946	419,863	869,809
257-C369-T21	766,056	347,397	279,540	626,937
258-C422-T19	994,866	416,125	372,219	788344
259-C304-T22	652,389	283,832	262,989	546,821
260-C89-T22	725,504	274,635	273,769	548,404
261-C1281-T25	723,929	295,287	288,187	583,474
97-C120-T20	272,215	111,677	93,477	205,154
Lab-L-R1	541,990	232,529	195,545	428,074
Lab-M-R1	505,407	914,52	76,575	168,027
Lab-S-R1	482,356	79,818	62,906	142,724
Mouse-R1	842,854	370,097	314,552	684,649

Table A.3: Sequencing of degenerate and MNase digest gRNA libraries - Read counts

A.4 Data analysis: Read tables for sequencing of gRNAs from sorted cells

Sample name	Number of sorted cells	Barcoding PCR cycles	Number of reads total	Trimmed gRNA reads (min. length = 2)	Reads uniquely aligned to gRNA library (no MM)	Reads uniquely aligned (2 MM)	Reads uniquely aligned (3MM)	Reads uniquely aligned (2MM 1 gap)	Reads 5' barcode after trimming (-m 7) and Q20 filtering	Reads 3' barcode after trimming (-m 7) and Q20 filtering	Reads 14 bp UMI and uniquely mapped gRNA
511-B5-Set7-C1-R1	3387	2	3462418	1481439	747359	1032032	1034115	1146208	1956750	1998699	972963
512-B5-Set7-C1-R2	6494	2	2749006	324378	140949	164770	165273	166511	544318	531166	161352
513-B5-Set7-C3-R1	60613	2	3434644	2816771	1663010	2104928	2114260	2213812	3092827	3059063	2048241
514-B5-Set7-C3-R2	60628	2	5163172	4314936	2331293	2974509	2982160	3110249	4767132	4713945	2920104
515-B5-Set7-C4-R1	60818	2	5427544	2023361	1212004	1434575	1469612	1490452	4708568	4674164	1369977
516-B5-Set7-C4-R2	60597	2	4533664	2717295	1523388	1823983	1857520	1982207	3313033	3272653	1825831
517-B5-EMT-R1	35054	2	5197136	3024146	1624491	2031611	2032888	2066267	4438609	4305511	1901701
518-B5-EMT-R2	31750	2	3581568	1615048	654428	999166	1009204	1082024	1835971	1820483	1052452
475-B4-EMT-R1	5264	2	3755671	3216616	1772238	2226096	2228035	2324790	3196006	3186726	2124894
476-B4-p300-C20-R1	2673	2	4553547	3456343	1759967	2124045	2143911	2278755	3453638	3432512	2114928
477-B4-p300-C19-R1	7163	2	4809906	4071418	2231478	2951845	2961877	3023430	4201691	4153258	2839087
478-B4-p300-C19-R2	10719	2	5756603	5128558	2882332	3670377	3749550	3814696	5117624	5073746	3478316
479-B4-p300-C19-R3	7025	2	4870002	4033909	2162708	2815093	2830446	3037692	4044503	4074981	2769917
480-B4-p300-C12-R1	5712	2	4923695	3050973	1635795	2090801	2158412	2148267	4064362	4046882	1964784
481-B4-p300-C12-R2	7029	2	5245170	3521975	1907793	2475035	2483988	2563743	3683413	3668607	2330817
482-B4-p300-C12-R3	7601	2	4745406	3700431	2069613	2516972	2532306	2763640	3710328	3691318	2479306
483-B4-p300-C20-R2	2859	2	3665419	2989699	1512091	2038408	2041442	2103537	3090138	2980029	1802770
489-B3-Set7-R1	26088	2	3448986	3143762	1543234	1635550	2263603	2442448	2825438	3098103	1980654
490-B3-Set7-R2	18129	2	3783550	979799	104222	455850	456717	663570	973360	977014	656809
491-B3-Set7-R3	10261	2	3169233	1260130	536860	698251	699147	770782	1255361	1254312	762776
492-B3-Set7LL-R1	26681	2	3595255	2674400	1312057	1950383	1956848	2057448	2662770	2328850	1753702
493-B3-Set7LL-R2	27117	2	4530650	3938280	2942807	3084249	3085655	3130240	3932557	3925781	3031198
494-B3-Set7LL-R3	26497	2	6182431	5383468	1545427	2237196	2238618	2363193	5321580	5344938	2324735

Table A.4: Screens using the EMT5000 gRNA library - Read counts for sequencing of gRNAs amplified from sorted cells

Sample name	Number of sorted cells	Barcoding PCR cycles	Number of reads total	Trimmed gRNA reads (min. length = 2)	Reads uniquely aligned to gRNA library (no MM)	Reads uniquely aligned (2 MM)	Reads uniquely aligned (3MM)	Reads uniquely aligned (2MM 1 gap)	Reads 5' barcode after trimming (-m 7) and Q20 filtering	Reads 3' barcode after trimming (-m 7) and Q20 filtering	Reads 14 bp UMI and uniquely mapped gRNA
495-B3-EMT-R1	27095	2	2695208	1498348	466299	988404	1019970	1132378	1489639	1480101	1092945
496-B3-EMT-R2	26709	2	2208726	1459384	519708	865462	866541	867579	1445138	1416402	820599
497-B3-VP64-R1	26769	2	6691886	4781184	2350842	2462105	2463145	2472512	4445865	4761633	2173475
498-B3-VP64-R2	27368	2	5924724	5743371	2573347	3155879	3184277	3263273	5678204	5678542	3146304
565-3-B8-EMT-R1	24489	3	69552	4670	2619	3353	3391	3555	4670	4690	3380
566-3-B8-EMT-R2	25157	3	81692	3438	1849	2320	2355	2449	3438	3423	2282
567-3-B8-p300-C12-R1	10000	3	350365	25384	15287	19417	19716	20166	25448	25464	19426
568-3-B8-p300-C12-R2	27128	3	446530	144825	89630	110465	111288	116118	145727	145074	111139
569-3-B8-p300-C12-R3	28698	3	36450	1383	629	783	795	816	1315	1336	709
570-3-B8-p300-C19-R1	32249	3	1408638	125643	74983	92670	93431	102332	123660	125018	96388
571-3-B8-p300-C19-R2	36933	3	1471111	666221	398543	496673	501159	522941	662877	662723	504031
572-3-B8-p300-C19-R3	33259	3	450666	33063	20357	24982	25183	26802	32874	32884	26036
573-3-B8-p300-C20-R1	39101	3	1783672	445482	265594	334404	337002	358061	452626	452373	338093
574-3-B8-p300-C20-R2	47876	3	2304157	1710984	1030432	1266872	1276966	1338120	1696818	1698562	1291505
575-3-B8-p300-C20-R3	50141	3	2888375	2244812	1370819	1697730	1710361	1785577	2233540	2237515	1725375
565-2	24489	3	93376	87353	53107	66548	66881	70672	88553	88399	68166
566-2	25157	3	13970	12538	7724	9046	9344	10015	12486	12485	9592
567-2	10000	3	297402	265438	159216	204081	207672	211972	266327	266074	203617
568-2	27128	3	3305756	3227728	2004939	2470938	2487878	2596036	3238950	3227897	2493667
569-2	28698	3	13528	11155	6750	8134	8187	8579	11057	11033	8132
570-2	32249	3	1736948	1726268	1024224	1270340	1280884	1406672	1700111	1717376	1324355
571-2	36933	3	4463165	4436753	2647981	3296682	3327058	3474476	4413935	4411391	3343112
572-2	33259	3	351555	338337	210026	256376	258598	275114	337372	336811	266897
573-2	39101	3	6088712	5753751	3435053	4327723	4361535	4642027	5882231	5859597	4379744
574-2	47876	3	4865267	4849323	2899377	3571231	3600362	3774071	4809522	4813264	3636826
575-2	50141	3	4599211	4570632	2788249	3452822	3478638	3632780	4546573	4554827	3507315

Table A.4: Screens using the EMT5000 gRNA library - Read counts for sequencing of gRNAs amplified from sorted cells - continued

Sample name	Number of sorted cells	Barcoding PCR cycles	Number of reads total	Trimmed gRNA reads (min. length = 2)	Reads uniquely aligned to gRNA library (no MM)	Reads uniquely aligned (2 MM)	Reads uniquely aligned (3MM)	Reads uniquely aligned (2MM 1 gap)	Reads 5' barcode after trimming (-m 7) and Q20 filtering	Reads 3' barcode after trimming (-m 7) and Q20 filtering	Reads 14 bp UMI and uniquely mapped gRNA
1-Set7-C4-R1	30000	3	2679658	820457	486122	607710	611822	641066	814319	824476	609092
2-AV2	15000	3	1687671	302179	165991	207105	208602	217350	299799	308302	206085
3-Set7-C4-R2	30000	3	7153419	4412613	2593830	3247926	3270738	3412421	4373377	4388929	3254208
4-AV2	15000	3	3496637	1085999	642518	814092	822066	855928	1077441	1091742	819665
5-AV1	15000	3	3649716	2330014	1352694	1685037	1697506	1764233	2311235	2318080	1692025
6-AV1	15000	3	3999650	1842804	770607	976493	985390	1022290	1830426	1840385	969385
7-Set7-C3-R1	12000	3	2423929	833694	399163	492094	496785	514106	827091	840121	488384
8-Set7-C3-R2	12000	3	2001796	415000	228217	288821	290377	303589	411862	419294	287372
9-Set7LL-C22-R1	13000	3	4162412	2350017	1382321	1757253	1767382	1848962	2332432	2327452	1760660
10-Set7LL-C22-R2	7000	3	2212272	79599	31085	38692	39174	40842	78735	95941	38667
11-Set7LL-C1-R2	8000	3	3918574	1483765	901435	1128633	1135971	1184957	1471577	1486777	1127198
12-Set7LL-C1-R1	17000	3	2396657	752383	438168	550149	552416	575327	746968	743399	547629
13-p300-C20-R1	25000	3	4722347	3001978	1791207	2254705	2266414	2362551	2979520	2987708	2253863
14-p300-C20-R2	22000	3	4559636	276023	160568	200493	202448	210647	273510	303177	201113
15-P300-C12-R1	7000	3	1845894	2864	1458	1823	1850	1892	2765	18393	1769
16-p300-C12-R2	8000	3	3244691	895748	479836	600394	603201	631105	889026	905317	596626
C-AV1	10000	3	5046466	3025887	1733158	2178224	2193238	2304794	2988996	2955646	2190339
D-AV2	10000	3	3077885	1369562	811662	1010133	1016690	1054393	1353914	1367597	1001427
E-PV1	12000	3	2879984	5103	3382	3637	3667	3664	4988	24021	3517
F-PV2	15000	3	3866567	1026361	596099	765615	772791	809966	1014606	1030757	769789
H-SV1	13000	3	4105484	923303	535818	667586	674036	708984	916102	933948	675900
I-SV2	15000	3	3826770	1826940	1061593	1341816	1356132	1414624	1814769	1821790	1357373
J-SVL2	8000	3	3537408	2162570	1237324	1545588	1550722	1617121	2143827	2145828	1532609
K-SVL1	10000	3	3412821	110334	59414	79934	80066	84868	108702	124512	299024
TOTAL			239133662								97944143

Table A.4: Screens using the EMT5000 gRNA library - Read counts for sequencing of gRNAs amplified from sorted cells - continued

Appendix B

Bioinformatic data analysis - Detailed methods

B.1 Availability of scripts

A copy of this documentation as well as the scripts and files that are referred to below can be found on github (`https://github.com/annakoe/AnalysisScripts`).

B.2 Degenerate gRNA libraries

B.2.1 Finding all gRNA targets in the repeat-masked human genome

The genomic target sites of gRNAs designed for use with the *S.pyogenes* CRISPR/-Cas system are of the general form $GN_{20}GG$. The presence of the G at the start of the motif is required for initiation of transcription from the U6 promoter, present on many gRNA expression vectors. This is followed by 19 random bases (N) that determine the gRNA target site - these bases comprise the gRNA protospacer. The protospacer sequence is followed by the protospacer-adjacent motif (PAM), which is not part of the gRNA sequence but is present in the genomic target sequence just 3' of the gRNA target site. For the *S.pyogenes* CRISPR/Cas system, the PAM has to be comprised of the bases NGG.

A gRNA library is a pooled collection of gRNAs. A random library has a complexity of 4^{19}, which corresponds to 10^{12} different sequences. The aim here is to reduce this complexity and maximise binding to the genome. To this end, I first identified the number of occurrences of the sequences GN20GG in the human genome. The Bioconductor (Bioconductor version 2.14) package BSgenome [113] was used to identify all sequences matching this pattern in the repeat-masked human genome in R (version 3.1.1). An R script named `find_all_gRNAs.R` contains the code and outputs a file containing chromosome coordinates and strand information for all hits, with the last column containing the pattern identifier. The script looks for the $GN_{20}GG$ by default (and outputs a file named `GN20GG_masked_allregions.txt`), unless another pattern is supplied by the user as follows:

```
./find_all_gRNAs.R -p [pattern, e.g. "GTACN"], -n [pattern-
   name, e.g. "my-pattern"], -o [outputfilename, e.g. "
   All_GTACN_hg19_RM.txt"]
```

The regions mapping to chromosomes 1-22 and the sex chromosomes were extracted from the outputfile `GN20GG_masked_allregions.txt` as follows:

```
grep -v 'random' GN20GG_masked_allregions.txt | grep -v '
   hap' | grep -v 'chrUn' | grep -v chrM >
   GN20GG_masked_autoXY.txt;

#remove the last column
awk '{print $1 "\t" $2 "\t" $3 "\t" $4}'
   GN20GG_masked_autoXY.txt > GN20GG_masked_autoXY.bed;
```

B.2.2 Identifying gRNAs that overlap with known promoters

In order to generate an annotation file containing promoter regions, the annotation file for known transcripts (human GRCh37-70) was downloaded from Ensembl (version70) and parsed through a script supplied by Gareth Wilson. This script can be found here: `https://github.com/regmgw1/regmgw1_scripts/blob/master/ensembl_scripts/transcript2promoter.pl`). This script extracts coordinates -1000 and +500 bp from the start of the transcript. From this file non-overlapping promoter sequences were derived by strand-specific merging using the Bedtools suite (v2.17.0) [114].

```
#strand-specific merging of promoter annotation file
mergeBed -s -i promoters.gff | awk '{print "chr"$1 "\t" $2
   "\t" $3 "\t" $4}' > promoters_merged.bed

#use Bedtools to find regions that overlap promoter
# regions with minimum of 1 bp
intersectBed -a GN20GG_masked_autoXY.bed -b
   yourpath2annotation_files/human_GRCh37_70/
   promoters_merged.bed -wa >
   GN20GG_masked_autoXY_promoters_merged;
```

Then, FASTA coordinates were retrieved using twoBitToFa [137], which is available from the UCSC website (`http://hgdownload.cse.ucsc.edu/admin/exe/`).

```
#Separate according to whether pattern is on plus or
#minus strand
grep '+' GN20GG_masked_autoXY_promoters_merged >
```

```
GN20GG_masked_autoXY_promoters_merged_PLUS;

grep '-' GN20GG_masked_autoXY_promoters_merged >
GN20GG_masked_autoXY_promoters_merged_MINUS

#make into a gff file
awk '{print $1 ":" ($2 - 1) "-" $3}'
   GN20GG_masked_autoXY_promoters_PLUS  >
GN20GG_masked_autoXY_promoters_PLUS.gff;

twoBitToFa yourpath2/human/GRCh37/hg19.2bit
GN20GG_masked_autoXY_promoters_merged_PLUS.fa
-seqList=GN20GG_masked_autoXY_promoters_merged_PLUS.gff

#repeat for file containing hits on the minus strand
```

The Python script `reverse_complement_fasta.py` was used to reverse-complement the FASTA sequences on the minus strand [159]. The script was invoked as follows:

```
python reverse_complement_fasta.py
   GN20GG_masked_autoXY_promoters_merged_MINUS.fa >
GN20GG_masked_autoXY_promoters_merged_MINUS_REVERSE_Complement
   .fa
```

The fasta files on the plus and minus strand were combined using the cat command, lowercase letters converted to uppercase using the seqret tool from EMBOSS [160], and sequences collapsed into a unique set with `fastx_collapser` [161] .

```
#combine the files
cat GN20GG_masked_autoXY_promoters_merged_PLUS.fa
GN20GG_masked_autoXY_promoters_merged_MINUS_REVERSE_Complement
   .fa > GN20GG_masked_autoXY_promoters_merged_TOTAL.fa

#convert to uppercase
seqret GN20GG_masked_autoXY_promoters_merged_TOTAL.fa
GN20GG_masked_autoXY_promoters_merged_TOTAL_UPPER.fa -
   sformat fasta -supper Y

#get unique fasta sequences
```

```
fastx_collapser <
   GN20GG_masked_autoXY_promoters_merged_TOTAL_UPPER.fa >
GN20GG_masked_autoXY_promoters_merged_PlusMinus_UNIQUE.fa
```

This yields a file containing 4,113,530 sequences.

B.2.3 Identifying a consensus sequence for gRNAs that fall into promoters

The Bioconductor package Biostrings [116] was used to derive a consensus sequence from the list of FASTA sequences generated above as follows (run in R):

```
>library(Biostrings)

>promMINUS<-readDNAStringSet(
"GN20GG_masked_autoXY_promoter_minus_REVERSECOMPLEMENT.fa",
    format="fasta")

>fm<-consensusMatrix(promMINUS)
minus<-fm[1:4,]
pwm_minus<-t(t(minus)/rowSums(t(minus)))
```

The following code was used to generate sequence logo plots [162] in R.

```
>library(ggplot2)

>berrylogo<-function(pwm,gc_content=0.5,zero=.0001){
 backFreq<-list(A=(1-gc_content)/2,C=gc_content/2,G=
    gc_content/2,T=
  (1-gc_content)/2)
  pwm[pwm==0]<-zero
bval<-plyr::laply(names(backFreq),function(x){log(pwm[x,])-
   log(
  backFreq[[x]])})
row.names(bval)<-names(backFreq)
p<-ggplot2::ggplot(reshape2::melt(bval,varnames=c("nt","pos
   ")),
  ggplot2::aes(x=pos,y=value,label=nt))+
    ggplot2::geom_abline(ggplot2::aes(slope=0), colour = "
       grey",size=2)+
```

```
    ggplot2::geom_text(ggplot2::aes(colour=factor(nt)),size
       =8)+
    ggplot2::theme(legend.position="none")+
    ggplot2::scale_x_continuous(name="Position",breaks=1:
       ncol(bval))+
    ggplot2::scale_y_continuous(name="Log relative
       frequency")
  return(p)
}

#invoke the function with:
berrylogo(pwm_minus, gc_content=0.5, zero=.0001)
```

B.2.4 Reducing complexity by identifying the most significant clusters

I reasoned that it might be possible to reduce the complexity of a random library by clustering. I defined regions of interest as the 4,671,728 genomic hits of the form GN20GG that fall into or next to known promoter sequences in the human genome with a minimum of 1 bp overlap. I used LCS-HIT (Version 0.5.2) [115] to cluster those regions of interest on the basis of sequence similarity. Clusters were ranked in descending order by the number of members and the top 15 clusters extracted.

```
#cluster sequences based on sequence similarity threshold
# of 0.2 and use the "exact algorithm"
lcs_hit-0.5.21/lcs_hit -i GN20GG_masked_autoXY_promoters
_merged_PlusMinus_UNIQUE.fa -O LCSHIT_OUTPUT20G1 -c 0.2 -g
   1 &
```

LCS-HIT outputs the FASTA-identifiers only, therefore FASTA sequences were extracted using the script `RetrieveFasta.pl`, downloaded from reference [163]. This step is exemplified here for the top cluster (Cluster190).

```
./RetrieveFasta.pl Cluster190
   GN20GG_masked_autoXY_promoters_merged
_PlusMinus_UNIQUE.fa > Cluster190.fa;
```

For each of the top 15 clusters consensus sequences were computed using the Bioconductor package Biostrings (run in R).

```
>library(Biostrings)
>Clust190<-readDNAStringSet("Cluster190.fa", format="fasta
   ")
>consensusString(Clust190, ambiguityMap=IUPAC_CODE_MAP,
   threshold=0.19,
shift=0L, width=NULL)
```

The threshold option allows the user to define the percentage threshold at which a given nucleotide will be incorporated into the consensus sequence at a given position. The threshold was varied between 0.14 and 0.25 so as to yield consensus sequences representing roughly 10^4, 10^5 and 10^6 different sequences respectively. The complexity, or number of sequences represented by a cluster, was computed using a C script downloaded from [164] and named `AllSequencesFromConsensus.c`.

```
#compile with
gcc -o AllSequencesFromConsensus AllSequencesFromConsensus.
   c

#run the script and pipe the output to line-count (as
# shown for the consensus sequence of Cluster190_T20)
./AllSequencesFromConsensus RRRGRGRRRRRGRRGRRGV | wc -l
```

Next, the "GenomeSearch" function of the Biostrings package (see Section B.2.1) was used to run each of the consensus sequences for each of the top 15 clusters against the masked human genome.

```
#store the sequences that should be run against the genome
# in a DNAStringSet object dict0, here shown for Cluster190
>dict0<-DNAStringSet(c("GRRRGRGRRRRRGRRGRRGVNGG",
"GRRRRRRRRRRRRRGRRGVNGG", "GVRRRRRRRRRRRRRRRRSVNGG",
 "GVRRRRRRRRRRRRRRRRSVNGG", "GVVRRRRRRRRRRRRRRVVVNGG", "
    GVVVRRRRRRRRRRRRVVVVNGG"))

#provide an identifier for each of the consensus sequences,
#here, cluster number and chosen threshold were used,
#again shown for Cluster190
>names(dict0)<-c("Cluster190_T20", "Cluster190_T19", "
   Cluster190_T18", "Cluster190_T17", "Cluster190_T16", "
   Cluster190_T15")

#invoke as follows:
```

```
>GenomeSearch_masked(dict0, outfile="
   Top15Clusters_masked_allregions.txt")
```

Because gRNAs are known to tolerate mismatches in their target sites [165], I reasoned that counting only exact matches probably underestimates the true number of target sites of any given sequence. Thus, the analysis was repeated allowing for an arbitrary number of up to 3 mismatches (gRNAs are known to tolerate more mismatches, especially towards the 5' end of the protospacer sequence. However, position of the mismatch in the gRNA was not taken into account here).

```
#allowing for a maximum of three mismatches
#change relevant parameter in the "GenomeSearch" function
#of the code

>plus_matches <- matchPattern(pattern, subject, max.
   mismatch=3, min.mismatch=0, fixed=c(pattern=FALSE,
   subject=TRUE))

>runAnalysis_masked_3mismatch(dict0, outfile="
   Top15Clusters_masked_allregions_3mismatch.txt")
```

Output files from both scripts were processed as follows and the results stored in Table 3.1 of the PhD thesis.

```
# retrieve hits for autosomes and sex chromosomes only
grep "Cluster190_T20" Top15Clusters_masked_allregions.txt
   | grep -v 'random' | grep -v 'hap' | grep -v 'chrUn' |
   grep -v chrM | awk '{print $1 "\t" $2 "\t" $3}' >
   Cluster190_T20_autoXY

#count number of hits that fall into promoter regions
intersectBed -a Cluster190_T20_autoXY -b annotation_files/
   promoters_merged.bed -wa -wb | sortBed | wc -l;

#count the number of unique promoter regions hit
intersectBed -a  Cluster190_T20_autoXY -b  annotation_files
   /promoters_merged.bed -wb | cut -f 5-8 | sortBed | uniq
   | wc -l;
```

I defined the targeting efficiency for each of the possible consensus sequences of a given cluster by dividing the number of promoter hits by the total number of unique sequences (complexity) of the consensus. For each cluster the consensus

sequences with the highest targeting sequence was chosen. The 15 top clusters were then ranked by targeting efficiency and the top 6 clusters chosen for library preparation.

B.3 Data analysis: sequencing of the degenerate and MNase digest libraries

B.3.1 Read trimming and quality filtering

Addition of sequencing adapters was non-directional, which means the reads contain gRNA cassettes both in forward and reverse direction.

First, the gRNAs that were sequenced in the forward direction were extracted using cutadapt [120] to trim away the plasmid sequence (shown here for the fastq sequencing data file generated from the Random GN19 library).

```
cutadapt -g
   AAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACC -O
   41 -m 2 --untrimmed-output=Untrimmed255_R1_50F.fastq.gz
   -o 255-N19_R1_F50_trimmed.fastq.gz ../255-
   N19_S5_L001_R1_001.fastq.gz;

cutadapt -a GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCG -O
   33 -m 2 --untrimmed-output=Untrimmed255_R1_19R.fastq.gz
   -o 255-N19_R1_F50+19R_trimmed.fastq.gz 255-
   N19_R1_F50_trimmed.fastq.gz;
```

Next, the reads stored in the untrimmed output, which will contain reads of the cassette sequenced in the reverse direction were concatenated, sorted and gRNA sequences extracted as follows (again showing example of reads from the GN19 Random library):

```
cat Untrimmed255_R1_50F.fastq.gz Untrimmed255_R1_19R.fastq.
   gz > Untrimmed255_R1_50F+19R.fastq.gz;

zcat Untrimmed255_R1_50F+19R.fastq.gz | paste - - - - |
   sort -k1,1 -t " " | tr "\t" "\n" > Untrimmed255_R1_50F
   +19R_sorted.fastq;
```

```
cutadapt -a
   GGTGTTTCGTCCTTTCCACAAGATATATAAAGCCAAGAAATCGAAATACTT -O
   41 -m 2 --untrimmed-output=Untrimmed255_50FRC.fastq.gz -
   o 255_R1_50FRC_trimmed.fastq.gz Untrimmed255_R1_50F+19
   R_sorted.fastq;

cutadapt -g CGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC -O
   33 -m 2 --untrimmed-output=Untrimmed255_19RRC.fastq.gz -
   o 255_R1_50FRC+19RRC_trimmed.fastq.gz 255
   _R1_50FRC_trimmed.fastq.gz;
```

This results in extraction of gRNA sequences sequenced in the reverse direction.

```
### Double untrimmed reads:
cat Untrimmed255_50FRC.fastq.gz Untrimmed255_19RRC.fastq.gz
    > Untrimmed255_50FRC+19RRC.fastq.gz;

zcat Untrimmed255_50FRC+19RRC.fastq.gz | paste - - - - |
   sort -k1,1 -t " " | tr "\t" "\n" > Untrimmed255_50FRC+19
   RRC_sorted.fastq;

### Successfully trimmed reads:

cat 255-N19_R1_F50+19R_trimmed.fastq.gz 255_R1_50FRC+19
   RRC_trimmed.fastq.gz > 255_R1_4waytrimmed.fastq.gz;

zcat 255_R1_4waytrimmed.fastq.gz | paste - - - - | sort -k1
   ,1 -t " " | tr "\t" "\n" > 255_R1_4waytrimmed_sorted.
   fastq;

### Extracted gRNA sequences sequenced in reverese
   direction need to be reverse-complemented

zcat 255_R1_50FRC+19RRC_trimmed.fastq.gz |
   fastx_reverse_complement -Q33 -z > 255_R1_50FRC+19
   RRC_trimmed_onedirection.fastq.gz;


cat 255-N19_R1_F50+19R_trimmed.fastq.gz 255_R1_50FRC+19
   RRC_trimmed_onedirection.fastq.gz > 255
   _R1_4waytrimmed_onedirection.fastq.gz;
```

```
zcat 255_R1_4waytrimmed_onedirection.fastq.gz | paste - - -
    - | sort -k1,1 -t " " | tr "\t" "\n" > 255
   _R1_4waytrimmed_sorted_onedirection.fastq;
```

Next, read length distribution between 15 and 40 bp were plotted for each of the libraries (again shown here for the file containing reads from the GN19 Random library):

```
awk '{y= i++ % 4 ; L[y]=$0; if(y==3 && length(L[1])<=40) {
   printf("%s\n%s\n%s\n%s\n",L[0],L[1],L[2],L[3]);}}' 255
   _R1_4waytrimmed_sorted.fastq > 255
   _R1_4waytrimmed_sorted_smaller40.fastq;

awk '{if(NR%4==2) print}' 255
   _R1_4waytrimmed_sorted_smaller40.fastq | awk '{print
   length($1)}' | sort -n >  255
   _R1_4waytrimmed_sorted_smaller40_lengths;
```

Per base sequence quality after trimming was assessed using FASTQC [121]

```
fastqc 255_R1_4waytrimmed_sorted_smaller40.fastq  --outdir
   =../FASTQC_Trimmedreads
```

B.3.2 Histogram of gRNA lengths

This code was used to generate the plots in Figure 3.12 of the PhD thesis. Histograms of read lengths were generated in R:

```
reads<-read.table("255
   _R1_4waytrimmed_sorted_smaller40_lengths", header=F)

par(mar=c(5,6+2,4,2)+1)
hist(reads[,1], breaks=seq(0,40,by=1), col="grey", main="
   Method 1 - Random N19", xlab="Insert length (bp)", ylim=
   c(0, 500000), las=1, ylab="", cex.main=1.3, cex.axis
   =1.3, cex.lab=1.3)
title(ylab = "Frequency", line = 6, cex.lab=1.3)
```

B.3.3 Histogram of read frequencies

This code was used to generate the plots in Figure 3.11 of the PhD thesis.

```
awk '{if(NR%4==2) print}' 255
   _R1_4waytrimmed_sorted_onedirection_smaller40.fastq |
   sort | uniq -c | sort | sed 's/^ *//' > 255
   _R1_4waytrimmed_sorted_onedirection_smaller40_sortedbyfrequency
   ;

#Histograms of read frequency were generated in R:
reads<-read.table("255
   _R1_4waytrimmed_sorted_onedirection_smaller40_sortedbyfrequency
   ", sep=" ", head=F)
colnames(reads)<-c("counts", "sequence")
reads_upto5<-reads[reads$counts<=5,]
reads_greater5<-reads[reads$counts>=5,]
par(oma=c(0,0,0,0))
par(mar=c(6,6,8,3))
par(fig=c(0.1,1,0,0.75))
plot(rownames(reads), reads$counts, type="p", xlab="gRNA
   ranked by counts", xlim= c(0, 600000), ylim=c(5,350),
   ylab="", las=1, col="orange", lwd=4, cex.axis=1.7, cex.
   lab=1.9)
title(ylab = "Read counts", line = 4, cex.lab=1.9)
par(fig=c(0.1,1,0.45,1), new=TRUE)
plot(rownames(reads_upto5), reads_upto5$counts, type="l",
   main="Method 3 - Mouse", xlab="", xlim= c(0, 600000),
   ylim=c(0,5), ylab="", lwd=8, col="orange", las=1, cex.
   axis=1.7, cex.lab=1.9, cex.main=1.9)
title(ylab = "Read counts", line = 4, cex.lab=1.9)
```

B.3.4 Alignment to the reference genome

Trimmed reads were aligned to the reference genome, either human (GRCh37/hg19) or mouse (NCBI37/mm9) as appropriate, using BWA [126], again shown using the file containing reads from the N19 Random library as an example.

```
#use bwa to align reads to human genome without allowing
   mismatches and printing all alignments, seed length is
   set to 19
```

```
bwa aln -n 0 -o 0 -l 19 -N /mnt/store2/local_data/
   genomic_data/human/GRCh37/human_GRCh37.tmp 255
   _R1_4waytrimmed_sorted.fastq > 255_R1_4waytrimmed.sai

#generate the sam file
bwa samse -n 10000 /path2/human_GRCh37.tmp 255
   _R1_4waytrimmed.sai 255_R1_4waytrimmed_sorted.fastq >
   255_R1_4waytrimmed.sam
```

Uniquely mapped reads were counted using Samtools [166].

```
samtools view -S -q1 255_R1_4waytrimmed.sam | wc -l
```

The remainder of the analysis (parts of which are shown in Fig. 3.18 on page 111 of the thesis) was conducted by my collaborator Karolina Worf at the Helmholtz Institute in Munich. The documentation for this analysis is available at hmgubox (`https://hmgubox.helmholtz-muenchen.de:8001/d/6c6e75236e/`; password: Coralina).

B.4 Design of the EMT5000 library

B.4.1 Defining regions of interest

A list of genes known to be involved in the regulation of epithelial-to-mesenchymal transition (EMT) was taken from DeCraene *et al.* [123] (**table 2.9**).

B.4.2 Identifying all gRNAs that fall into regions of interest

The genomic regions listed in **Table 2.9** where chosen as target sites for the design of gRNAs and saved in the file `EMT_genepromoter_comprehensive.gff`. All gRNAs falling into these regions were then found using the Bedtools [114] intersect function on the file containing all gRNAs with an NGG PAM found in the genome (see section B.2.1 for how this file was generated).

```
intersectBed -a GN20GG_masked_autoXY.gff -b
   EMT_genepromoter_comprehensive.gff -f 1 -wa -wb >
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive.gff
```

B.4.3 Subsetting only gRNAs that map uniquely to the human genome

Guide RNAs were then aligned to the genome in order to identify those that map uniquely. Before alignment, files had to be converted into the correct format as follows:

```
##for alignment need to first convert file into fasta
   format
##separate + and - strand
cut -f 1,2,3,4
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive.gff
   | grep '+' | awk '{print "chr" $1 ":" ($2-1) "-" $3}' >
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_PLUS
   .2bit;

cut -f 1,2,3,4
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive.gff
   | grep -v '+' | awk '{print "chr" $1 ":" $2 "-" $3}' >
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_MINUS
   .2bit;

##convert to FASTA format using twoBitToFa
twoBitToFa /path2/human/GRCh37/hg19.2bit
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_PLUS
   .fa seqList=
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_PLUS
   .2bit;

twoBitToFa  /path2/human/GRCh37/hg19.2bit
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_MINUS
   .fa seqList=
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_MINUS
   .2bit;

##create reverse complement of guide RNAs on the minus
##strand using the script reverse_complement_fasta.py

python reverse_complement_fasta.py
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_MINUS
```

```
   .fa >
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_MINUS_RC
   .fa;

##combine the files for the + and - strand
cat
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_PLUS
   .fa
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_MINUS_RC
   .fa >
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive.fa;
```

The PAM sequence was removed using `trimming.py` and alignment to the genome without the PAM (i.e. as GN19 instead of GN20GG).

```
bwa aln -n 0 -o 0 -l 10 -N -I ~/path_to/GRCh37/human_GRCh37
   .tmp
GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_noPAM.
   fa >
GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_noPAM.
   sai;

bwa samse -n 10000 ~/path_to/GRCh37/human_GRCh37.tmp
GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_noPAM.
   sai
GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_noPAM.
   fa >
GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_noPAM.
   sam;

samtools view -S -q1
   GN20GG_masked_autoXY_EMT_genepromoter_
comprehensive_complete_noPAM.sam | cut -f  1 | awk -F ':'
   '{print $1 "\t" $2}' | awk -F '-' '{print $1 "\t" $2 "\t
   " $3}' | sortBed | uniq > GN20GG_masked_autoXY_EMT_
genepromoter_comprehensive_complete_noPAM_unique.bed
```

This file contains 5086 sequences, i.e. 5086 gRNA sequences of the form GN19 that align uniquely to the genome and are followed by an NGG PAM and are found in the regions of interest.

B.4.4 Addition of plasmid sequencs for use in Gibson cloning

For cloning into the gRNA vector pgRNA-pLKO.1 (see Methods) by Gibson cloning, vector-derived sequences were attached to either end of the gRNA sequence. To this end the FASTA files for the final set of 5086 unique gRNAs were retrieved and sequences appended as follows:

```
#split according to + and - strand

grep '+'
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_complete_
    noPAM_unique_strand.bed | awk '{print $1 ":" $2 "-" $3
   }' >
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_complete_
    noPAM_unique_strand_PLUS.2bit

grep -v '+'
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_complete_
    noPAM_unique_strand.bed | awk '{print $1 ":" $2 "-" $3
   }' >
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_complete_
    noPAM_unique_strand_MINUS.2bit

#convert to FASTA format and reverse-complement sequences
   on the
#minus strand

twoBitToFa /path2/human/GRCh37/hg19.2bit
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_complete_
    noPAM_unique_strand_PLUS.fa seqList=
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_complete_
    noPAM_unique_strand_PLUS.2bit;

twoBitToFa /path2/human/GRCh37/hg19.2bit
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_complete_
    noPAM_unique_strand_MINUS.fa seqList=
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_complete_
    noPAM_unique_strand_MINUS.2bit

python reverse_complement_fasta.py
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_complete_
```

```
    noPAM_unique_strand_MINUS.fa >
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_complete_
    noPAM_unique_strand_MINUS_RC.fa

#merge the two files

cat
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_complete_

noPAM_unique_strand_PLUS.fa
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_complete_

noPAM_unique_strand_MINUS_RC.fa >
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_complete_

noPAM_unique_strand.fa

# use trimming.py to remove PAM (because coordinates
   extracted
# from SAM file contained PAM, while alignment was run
   without
# PAM) and paste the vector sequences for Gibson cloning on
# either side of the gRNA sequence

sed 'n; s/$/GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCT/'
   GN20GG_masked_autoXY_EMT_genepromoter_
   comprehensive_complete_noPAM_unique.fa | sed 'n; s/^/
   TCTTGTGGAAAGGACGAAACACC/g' | paste - - | awk '{print $2
   "\t" $1}' > EMT_guides_Custom_Array.txt
```

A pool of sequences of the form 5'-TCTTGTGGAAAGGACGAAACACC-GN19-GTTTTAGAGCT AGAAATAGCAAGTTAAAATAAGGCT-3', where GN19 donates the 5086 different guide sequences was then ordered from Custom Array Inc.

B.5 Data analysis: Screens with the EMT5000 library and stable cell lines expressing the dCas9-SET7 or dCas9-p300 chromatin modifier

B.5.1 Read trimming and Quality Filtering

gRNA sequences integrated into the genome of FACS-sorted cells were amplified using PCR to add Illumina Nextera adapters. Libraries were sequenced on the Illumina HiSeq instrument. The resulting sequencing reads have the following general structure:

NNNNNNNAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCG-N19-GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGNNNNNNN

The first 7 N bases are the 5' barcode, followed by the plasmid 'stuffer'. GN19 denotes the gRNA sequence from the EMT5000 library, which is followed by another plasmid sequence and the last 7 N bases are the 3' barcode. The 5' and 3' barcodes serve as unique molecular identifiers (UMIs), allowing counting of original gRNA sequences extracted from lentivirus-infected cells by removing PCR-amplification bias.

Sequences from Lane1 and Lane2 of the flow cell were combined using the unix 'cat' command. Next, the 5' barcode was extracted from the reads using cutadapt (version 1.2.1) [120], requiring a minimum overlap of 35 bp between the plasmid stuffer sequence and the read with a maximum error of 10 % and a minimum barcode length of 7 bp.

```
Command line parameters:
-a AAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACC -O
   35 -m 7

# example bash loop to run cutadapt over all fastq files
# in the directory
for i in *.fastq.gz
do cutadapt -a
   AAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACC -O
   35 -m 7 --untrimmed-output="${i%.fastq.gz}"
   _5bc_untrimmed.fastq.gz -o "${i%.fastq.gz}"_5bc.fastq.gz
```

```
    "$i";
done
```

The 3' barcode was retrieved from the read in an analogous way:

```
Command line parameters:
-g GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCG -O 28 -m 7

# example bash loop to run cutadapt over all fastq files
# in the directory
for i in *.fastq.gz
do cutadapt -g GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCG -
   O 28 -m 7 --untrimmed-output="${i%.fastq.gz}"
   _3bc_untrimmed.fastq.gz -o "${i%.fastq.gz}"_3bc.fastq.gz
    "$i";
done
```

Finally, the gRNA sequence was extracted from the read, requiring a minimum length of 2 bp (to discard reads that contain no gRNAs and are derived from primer dimers):

```
Command line parameters:
cutadapt -g
   AAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACC -a
   GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCG -n 2 -m 2 -O
   10

# example bash loop to run cutadapt over all fastq files
# in the directory
for i in *.fastq.gz
do cutadapt -g
   AAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACC -a
   GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCG -n 2 -m 2 -O
   10 --untrimmed-output="${i%.fastq.gz}"_gRNA_untrimmed.
   fastq.gz -o "${i%.fastq.gz}"_gRNA.fastq.gz "$i";
done
```

Next, the barcode and gRNA reads were quality-filtered using `fastq_quality_filter` from the fastx-toolbox [161]. Reads where any base has a Quality score of less than 20 were discarded (Example code below is for the gRNA reads.)

```
Command line parameters: -Q33 -q 20 -p 1
```

```
#loop over all files in directory:
for i in *_gRNA.fastq; do fastq_quality_filter -Q33 -q 20 -
   p 1 -i "$i" -o "${i%_gRNA.fastq}"_gRNA_Q20.fastq; done
```

Subsequently files were converted from fastq to fasta format using `fastq_to_fasta` from the fastx-toolbox [161].

```
for i in *_gRNA_Q20.fastq;
do fastq_to_fasta -Q33 -n -i  "$i" -o "${i%gRNA_Q20.fastq}"
   gRNA_Q20.fasta;
done
```

B.5.2 Alignment to the EMT5000 library

Guide RNA reads were next aligned back onto the indexed EMT5000 reference library using bwa (version: 0.6.2-r126) [126].

The indexed EMT5000 library file used as a reference for alignment and was derived from the file `GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_complete_noPAM_unique_strand_PAMremoved.fa` (see section B.4.1) using `bwa index`:

```
bwa index
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_
   complete_noPAM_unique_strand_PAMremoved.fa
```

I empirically tested the following alignment parameters:

```
#Command line options no mismatches, no indels:
bwa aln -n 0 -o 0 -l 5 -N -I

#Command line options 2 mismatches, no indels:
bwa aln -n 2 -o 0 -l 5 -N -I

#Command line options 3 mismatches, no indels:
bwa aln -n 3 -o 0 -l 5 -N -I

#Command line options 2 mismatches and default open gaps
   (1):
bwa aln -n 2 -l 5 -N -I
```

```
#Command line options 3 mismatches and default open gaps
   (1):
do bwa aln -n 3 -l 5 -N -I
```

I found that allowing 2 mismatches and 1 gap gave the best alignment (see also the table of read numbers included in Appendix A). The alignment was invoked as follows:

```
for i in ../*_Q20.fasta;
do bwa aln -n 2 -l 5 -N -I EMT5000_library_reference/
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_
complete _noPAM_unique_strand_PAMremoved.fa "$i" > "${i%Q20
   .fasta}"Q20_aligned_2mismatches1gap.sai;
done

for i in *Q20_aligned_2mismatches1gap.sai;
do bwa samse -n 10000 EMT5000_library_reference/
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_
complete_noPAM_unique_strand_PAMremoved.fa "$i" "${i%
   Q20_aligned_2mismatches1gap.sai}"Q20.fasta > "${i%
   Q20_aligned_2mismatches1gap.sai}"
   Q20_aligned_2mismatches1gap.sam;
done
```

Following alignment allowing 2 mismatches and 1 gap, reads that mapped uniquely to the forward strand were extracted using Samtools [166]:

```
samtools view -F 20 -q 1 -S aligned_gRNA.sam  >
   gRNA_uniquely_mapped.sam
```

To check how many different gRNAs from the library were sequenced in each sample, use:

```
samtools view -F 20 -S  aligned_gRNA.sam | cut -f 3 | sort
   | uniq | wc -l
```

B.5.3 Combining the gRNA and adapter sequences in a single file

For subsequent analysis it was necessary to construct a tab-separated file with 3 columns containing the FASTA identifier (read-ID), gRNA chr:start-end and

barcode sequence for each read.

```
# extract the gRNA:
samtools view -F 20 -q 1 -S aligned_gRNA.sam | awk '{
   print$1 ".\t" $3}'  > gRNA_uniquelymapped

# extract the FASTA identifier:
samtools view -F 20 -q 1 -S aligned_gRNA.sam | awk '{
   print$1 "."}' > gRNA_uniquelymapped_readID
```

The files containing the quality-filtered 5' and 3' barcode (UMI sequence) generated above, were modified as follows (shown for 5' barcode only):

```
for i in *_5bc_Q20.fasta
do cat "$i" | paste - - | awk -F ' ' '{print $1 ".\t" $3}'
   > "${i%_5bc_Q20.fasta}"_5bc_Q20_point.fasta;
done
```

Next, only the barcodes associated with gRNAs that aligned uniquely were retrieved using grep:

```
for i in *_5bc_Q20_point.fasta
do grep -wFf "${i%_5bc_Q20_point.fasta}"
   gRNA_uniquelymapped_readID "$i" > "${i%_5bc_Q20_point.
   fasta}"gRNA_uniquelymapped_readID_with_5bc;
done
```

For each read, identified by its readID, the gRNA sequence and 5' and 3' barcodes were combined into a single file. The three files were joined (finding the union) using the JOIN command, which requires the files to be sorted in the following way:

```
for i in *gRNA_uniquelymapped;
do awk -F "\t" '{print ">" $1 "\t" $2}' "$i"| sort -k 1b,1
   > "${i%}"_sorted;
done

for i in *gRNA_uniquelymapped_readID_with_3bc;
do sort -k 1b,1 "$i" > "${i%}"_sorted;
done

for i in *gRNA_uniquelymapped_readID_with_5bc; do sort -k 1
   b,1 "$i" > "${i%}"_sorted;
```

```
done
```

Next, the three files were joined as follows:

```
for i in *gRNA_uniquelymapped_sorted;
do join "$i" "${i%gRNA_uniquelymapped_sorted}"
   gRNA_uniquelymapped_readID_with_5bc_sorted | join - "${i
   %gRNA_uniquelymapped_sorted}"
   gRNA_uniquelymapped_readID_with_3bc_sorted > "${i%
   gRNA_uniquelymapped_sorted}"
   gRNA_uniquelymapped_readID_gRNA_5bc_3bc_length14;
done
```

This yields a file containing for each uniquely mapped read its readID, the gRNA it mapped to (chr:start-stop) and the UMI found in the read (of length exactly 14 bp).

B.6 Deriving gRNA counts from UMI-barcodes without PCR error correction

The gRNA counts were derived by counting the number of times each gRNA occurs together with each barcode, which acts as a unique molecular identifier (UMI). To do the gRNA counting without any error correction, the script `collapse_barcodes.py` was run using a maximum edit distance of 0, i.e. not correcting any PCR errors that might have occurred in the barcode. This script can be invoked as follows:

```
python collapse_barcodes.py Inputfilename 0
```

The script accepts a tab-separated file with columns for (1) read-ID, (2) gRNA (chr:start-stop), (3) barcode and outputs two outputfiles with extension `_frequency_raw` and `frequency_no_orphans`. In the latter output orphan barcodes, i.e. barcodes that are only present in a single read, were removed prior to gRNA counting. Each output file has two comma-separated columns listing (1) the gRNA (chr:start-stop) and (2) the gRNA count.

The script was run over all files with the extension

`_uniquelymapped_readID_gRNA_5bc_3bc_length14` as follows:

```
for i in *length_14;
do python collapse_barcodes.py "$i" 0;
done
```

Assessing PCR error by plotting the number of reads per gRNA against number of different gRNA sequences

To assess whether PCR error drives barcode diversity, I treated the gRNA part of the read like a barcode and plotted the correlation between number of reads and counts (see Figure 4.8. on page 134 of the PhD thesis and also section B.6.3 below). This assumes that the likelihood of introducing an error into the sequence is the same for the UMI barcodes and gRNA portions of the amplicon.

The barcode sequence was replaced with the gRNA sequence for each read to generate a tab-separated file with columns (1) readID (2) gRNA chr:start-stop (3) 'pseudo-barcode' (gRNA sequence) as follows:

```
#get gRNA sequence
cat Sample_gRNA_Q20.fasta | paste - - | awk -F ' ' '{print
   $1 ".\t" $3}' | sort > Sample_gRNA_Q20_point

#get the read ID
awk -F '\t' '{print $1}' Sample_gRNA_5bc_3bc_length14 |
   sort > Sample_gRNA_5bc_3bc_length14_read_ID

#get gRNA sequence for each readID
grep -wFf Sample_gRNA_5bc_3bc_length14_read_ID
   Sample_gRNA_Q20_point >
   Sample_gRNA_Q20_point_uniquely_mapped

#sort the previously generate file containing [0] readID,
   [1] gRNA chr:start-stop, [2] UMI of 5' and 3' barcode
sort -k 1b,1
   Sample_gRNA_uniquelymapped_readID_gRNA_5bc_3bc_length14
   > Sample_gRNA_uniquelymapped_
readID_gRNA_5bc_3bc_length14_sorted

#sort the gRNA sequence file
sort -k 1b,1 Sample_gRNA_Q20_point_uniquely_mapped >
   Sample_gRNA_Q20_point_uniquely_mapped_sorted
```

```
#join the two files on readID
join
   Sample_gRNA_Q20_aligned_2mismatches1gap_uniquelymapped_
readID_gRNA_5bc_3bc_length14_sorted
   Sample_gRNA_Q20_point_uniquely_mapped_sorted | awk -F '
   ' '{print $1 "\t" $2 "\t" $4}' >
   Sample_gRNA_Q20_aligned_2mismatches1gap_uniquelymapped_
readID_gRNA_instead_of_barcode
```

This file was then run through `collapse_barcodes.py` as above and results were plotted as described in section B.6.3.

```
python collapse_barcodes.py
   Sample_gRNA_Q20_aligned_2mismatches1gap_
uniquelymapped_readID_gRNA_instead_of_barcode 0
```

B.6.1 Deriving gRNA counts from UMI-barcodes with naive PCR error correction

The script `collapse_barcodes.py` was run using a maximum edit distance of 4.

```
python collapse_barcodes.py Samplefilename 4
```

This means that before counting how many different barcodes are associated with each gRNA, the barcodes are collapsed into groups. Barcodes are first ranked in decreasing order based on the number of reads harbouring its sequence. The barcode with the most reads forms the first group. If the second-ranked barcode is within 4 edit-distances of this barcode it will be assumed to have originated by PCR error and will be added to the group. If the barcode differs from the group by greater than 4 edits, it will form its own group and so on. The number of groups per gRNA is the count after error correction. This reflects the number of original gRNA-barcode combinations.

The script was run over all files with the extension

`_uniquelymapped_readID_gRNA_5bc_3bc_length14` as follows:

```
for i in *length_14;
do python collapse_barcodes.py "$i" 4;
done
```

B.6.2 Bayesian PCR error correction of barcoded sequencing data

A Bayesian error correction script was written by James E. Barrett to infer gRNA counts from the UMI data. The model takes into account the fact that the 14 bp UMI consists of a 5' and 3' barcode that was attached to the gRNA amplicon during and initial round off PCR amplification during the sequencing library prep. The model infers the most likely number of initial gRNA-barcode data given the barcode sequences observed in the sequencing sample.

This Bayesian model takes as input the number of reads associated with each gRNA-UMI combination (without PCR error correction). I calculated these using the script make-csv-4Bayes.py. The script was run over all samples as follows:

```
for i in *length14; do python ../make_csv_4Bayes.py "$i";
   done
```

This script again takes the tab-separated three column file that lists (1) read ID, (2) gRNA (chr:start-end) and (3) barcode consisting of 5' UMI and 3' UMI fused together as input. The output is a csv file with three columns containing (1) gRNA (chr:start-end), (2) barcode consisting of 5' UMI and 3' UMI fused together and (3) number of reads associated with each barcode. The outputfile has the extension `_barcode_readcounts.csv` and is fed into a Bayesian error correction script described below.

Bayesian PCR error correction of barcoded sequencing count data script (by James E. Barrett)

This script and documentation was kindly contributed by James E. Barrett.

The Bayesian model infers the number of unique original barcoded gRNA molecules from noise-corrupted count data. The model estimates a corrected read count, which may be interpreted as a proxy for the original noise-free number of unique barcodes associated with a particular gRNA.

Model definition

For each gRNA we observe N barcode pairs denoted by $(\mathbf{y}_i^1, \mathbf{y}_i^2)$ where the superscript denotes the first and second barcodes and $i = 1, \ldots, N$. Elements of

the d-dimensional vector $\mathbf{y}_i^\eta \in \{\mathtt{T}, \mathtt{C}, \mathtt{G}, \mathtt{A}\}^d$ where $\eta = [1, 2]$. The number of corresponding sequencing reads is denoted by $\sigma_i \in \mathbb{Z}_+$.

The model assumes that there exist Q *latent barcodes* $\mathbf{x}_1^\eta, \ldots, \mathbf{x}_Q^\eta$ from which the observed barcodes are generated in a noise corrupting stochastic process (PCR amplification errors and random barcode switching). The model further assumes that for each pair $(\mathbf{y}_i^1, \mathbf{y}_i^2)$ only one of the observed barcodes is written in terms of the latent barcode via

$$\mathbf{y}_i^\eta = \sum_{q=1}^{Q} w_{iq}^\eta \theta(\mathbf{x}_q^\eta) \quad \text{subject to} \quad w_{iq}^\eta \in [0, 1] \quad \text{and} \quad \sum_{q,\eta} w_{iq}^\eta = 1. \tag{B.1}$$

There is therefore only one non-zero value of $[\mathbf{w}_i^1, \mathbf{w}_i^2]$ that indicates which latent barcode the observed pair is associated with. The function θ represents a noise corrupting stochastic process where the status of each nucleotide site may be changed randomly with probability $\beta \in [0, 1/2]$. We can therefore write

$$p(y_{i\mu}^\eta | x_{q\mu}^\eta, \beta) = \begin{cases} (1-\beta)\delta_{y_{i\mu}^\eta x_{q\mu}^\eta} + \beta(1 - \delta_{y_{i\mu}^\eta x_{q\mu}^\eta}) & \text{if } w_{iq}^\eta = 1 \\ 0 & \text{otherwise} \end{cases} \tag{B.2}$$

for $\mu = 1, \ldots, d$. We denote the collections of $\mathbf{x}_q^\eta$, $\mathbf{y}_i^\eta$ and $\mathbf{w}_i^\eta$ by $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{W}$ respectively. The posterior is

$$p(\mathbf{X}, \mathbf{W} | \mathbf{Y}, \boldsymbol{\sigma}, \beta) \propto p(\mathbf{Y} | \mathbf{X}, \mathbf{W}, \boldsymbol{\sigma}, \beta) p(\mathbf{X}) p(\mathbf{W}) \tag{B.3}$$

with

$$p(\mathbf{Y} | \mathbf{X}, \mathbf{W}, \boldsymbol{\sigma}, \beta) = \prod_i \left[\sum_{q,\eta} w_{iq}^\eta p(\mathbf{y}_i^\eta | \mathbf{x}_q^\eta, \beta) \right]^{\sigma_i}. \tag{B.4}$$

Maximum entropy priors for $\mathbf{X}$ and $\mathbf{W}$ are uniform distributions so $p(\mathbf{X})$ and $p(\mathbf{W})$ are constant.

Inference of model parameters

The Maximum A Posteriori (MAP) solution of $\mathbf{W}$ is denoted by $\mathbf{W}^*$. Since only one element of $[\mathbf{w}_i^1, \mathbf{w}_i^2]$ is non-zero the expression (B.4) is maximised by selecting $\text{argmax}_{q,\eta} p(\mathbf{y}_i^\eta | \mathbf{x}_q^\eta, \beta)$ as the non-zero element.

To find the MAP solution for nucleotide μ of the latent barcode indexed by (q, η) we consider all observed barcodes that generated from it (as defined by

$\mathbf{W}$). If we let n_1 and n_0 denote the total number of matches and mismatches respectively between that latent barcode and the associated observed barcodes, then the corresponding data likelihood is $(1-\beta)^{n^1_{q\mu}}\beta^{n^0_{q\mu}}$. This will be maximised if the number of matches is maximised. This is achieved selecting the most common observed nucleotide as the value for the latent nucleotide (while taking into account multiple counts).

If we let N_1 and N_0 denote the total number of matches and mismatches respectively across all of the latent barcodes and observed data then we can write

$$\log p(\mathbf{Y}|\mathbf{X}, \mathbf{W}, \beta) = N_1 \log(1-\beta) + N_0 \log \beta. \tag{B.5}$$

It is straightforward to show that the MAP estimate for beta is

$$\beta = \frac{N_0}{N_0 + N_1}. \tag{B.6}$$

The optimisation subroutine is initialised as follows:

Cluster into Q groups based on the *Hamming distance* between two barcodes (the Hamming distance is equivalent to the *edit distance*):

$$h(\mathbf{y}_i, \mathbf{y}_j) = \frac{1}{d}\sum_{\mu=1}^{d} \delta_{(1-y_{i\mu})y_{j\mu}}. \tag{B.7}$$

The corrected read counts are inferred as follows:

For a given value of Q we denote the value of the likelihood (B.4) at the MAP parameter estimate by

$$L(Q) = p(\mathbf{Y}|\mathbf{X}^*, \mathbf{W}^*, \beta^*). \tag{B.8}$$

The *Bayes information criterion* (BIC) score is defined by

$$\mathrm{BIC}(Q) = -2\log L(Q) + 2dQ \log N \tag{B.9}$$

where $2dQ$ is the number of free parameters in the model. The *corrected read count* is defined by

$$Q^* = \mathrm{argmin}_Q \mathrm{BIC}(Q). \tag{B.10}$$

Bayesian error correction script: The Code

The analysis is performed in R:

```
library(reshape2)

### Load and prepare a data file

# Length of barcode
D <- 7
# Load up one of the data files (needs to be in the current
    directory)
data <- read.csv("Samplename_length14_barcode_readcounts.
   csv",header=FALSE)
# vector of all the unique gRNA names
gRNA <- unique(data$V1)
# Total number of unique gRNAs
G <- length(gRNA)

### Generate datasets of barcodes

# Preallocate a list structure to hold the barcode datasets
Y <- vector('list',G)

# This loop goes through each gRNA, pulls out all the
   associated barcodes and puts them in a character matrix
for(mu in 1:G){
   ind <- which(data[[1]]==gRNA[mu])
   N <- length(ind)
   # Converts into character matrix (not the most elegant
      way...)
   Y[[mu]] <- matrix(as.vector(melt(lapply(as.character(
      data[[2]][ind]),strsplit,split=""))$value),nrow=N,
      ncol=2*D,byrow=TRUE)
}

### Fit model for each gRNA

# Preallocate a list of model resuls
res <- vector('list',G)
```

```
# Loop through gRNAs, for each one fit a model and get the
   corrected read count
# This can be parallelised for speed
for(mu in 1:G){
   # Begin tryCatch (catches any errors instead of stopping
       the loop)
   tryCatch({
      # Indices for barcodes matched that the current gRNA
      ind <- which(data[[1]]==gRNA[mu])
      # Vector of read counts
      counts <- data[[3]][ind]
      # Fit the model
      res[[mu]] <- fit_model(Y[[mu]], counts)
   }, error=function(e) NULL) #End tryCatch
} # End loop over gRNAs
```

This analysis calls functions stored in the R scripts `fit_model.R`, `LL.R` and `hamming.R`. A csv file of counts per gRNA for each sample (`bayesian_corrected.csv`) was then exported.

B.6.3 Diagnostic plot: Number of UMI-corrected counts versus number of reads per gRNA

Calculating the number of reads per gRNA

To calculate the number of reads per gRNA for each sample, I wrote the script `reads_per_gRNA.py`. This takes a 3 column tab-separated inputfile with the following columns: (1) read ID (2) gRNA chr:start-stop (3) 14 bp barcode and outputs a csv file with two columns: (1) gRNA chr:start-stop, (2) number of reads. The script can be invoked as follows:

```
for i in *length14; do python ./reads_per_gRNA.py "$i";
   done
```

Wrapping gRNA counts of all samples into a table

While the Bayesian model ouputs a csv file containing the counts per gRNA for each sample directly, the output of the script `collapse-barcodes.py` (used to

count gRNAs without error correction or to perform a naive PCR error correction) outputs one table per sample. This data can be merged into a single table listing for each gRNA the count in each sample using the script `make_table_from_counts.py`. This script also adds information about which gene is targeted by each gRNA. The script takes a variable number of inputfiles to wrap into a table:

```
python makeTable_from_counts.py INPUTFILE1 INPUTFILE2 ...
   INPUTFILEn
```

This was used to generate the tables `Dataframe_allsamples_readcounts.txt`

Generating the plots of counts versus number of reads for each gRNA

The number of reads per gRNA were plotted against the counts per gRNA, derived either without error correction, with naive PCR error correction or Bayesian PCR error correction (as described above), using the script `plot_counts_vs_number_of_reads.py`. Samplesheets can be found in the folder "samplefiles"

```
python plot_counts_vs_number_of_reads.py [Dataframe-Counts]
    [Dataframe-NumberOfReads] [Samplesheet]
```

```
# no PCR error correction
plot_counts_vs_number_of_reads.py
   Dataframe_allsamples_no_errors.txt
   Dataframe_allsamples_readcounts.txt samplefile.txt

# naive PCR error correction (4 mismatches)
plot_counts_vs_number_of_reads.py Dataframe_allsamples.txt
   Dataframe_allsamples_readcounts.txt samplefile.txt

# Bayesian error correction
plot_counts_vs_number_of_reads.py bayesian_corrected.csv
   Dataframe_allsamples_readcounts.txt samplefile.txt
```

This script was used to generate the plots in Figure 4.9 and Figure 4.10 of the PhD thesis.

B.6.4 Diagnostic plot: Counts versus number of sorted cells

This script accepts three user-supplied arguments, a dataframe of gRNA counts (c), a dataframe of numbers of sorted cells per sample (n), and a samplesheet (s) and returns a scatterplot of sorted cells versus counts with one data point per sample in the samplesheet. It is further hardcoded to color the dots according to whether two or three initial cycles of barcoding PCR were used to attach unique molecular identifiers to gRNA sequences before amplification and sequencing. The samplesheets can be found in the folder `samplefiles` and the file recording the number of sorted cells is in the folder `additional-files`. The script was invoked as follows:

```
python cellnumber_vs_counts.py -c bayesian_corrected.csv -n
    Number_of_sorted_cells.csv  -s samplefile_all.txt
```

This script was used to produce the graph in Figure 4.11 of the PhD thesis.

B.6.5 Enrichment analysis using DESeq2

Enrichment analysis was carried out using the DESeq2 package [127].

Preparing Bayesian error correction output for DESeq2

DESeq2 requires a list of counts per gRNA for each experiment to be analysed. The data for each experiment were extracted from the Bayesian analysis output file `bayesian_corrected.csv`. gRNAs with a lot of missing data where 3/4 of the counts in a given experiment are 0 (or NA) are removed before enrichment analysis using the script `make_input_4_DESeq.py` The samplesheets can be found in the folder `samplefiles`.

```
python make_input_4_DESeq.py -i bayesian_corrected.csv -s
   Samplefile_Batch8.txt
```

Using the above example, the script outputs a file named `Batch8_p300_counts.csv`.

Running DESeq2

This analysis was conducted using the script `DESeq2_script.R`. An example of how this script is run over the example file for experiment `Batch8_p300` is shown

below. DESeq2 further requires a file containing experimental information, specifying for each sample whether it belongs to the treatment or control group. These files can be found in the folder `additional_files/ColData_forDESeq2`.

```
./DESeq2_script.R -c path_to/DeSeq2/Inputfiles/
   Batch8_p300_counts.csv -e path_to/additional_files/
   ColData_forDESeq2/Batch8_DeSeq2_ColData -o '
   Batch8_p300_DESeq2_table.csv'
```

This outputs a csv file with the samplename and the extension `_DESeq2_table.csv`.

Visualizing enrichment

Log2 Fold Change between samples and controls for each gRNA was calculated using DESeq2 as described above. Log2Fold Change values were extracted from the output of DESeq2 (filename is `Batch8_p300_DESeq2_table.csv` for this particular example) and plotted along the chromosome for each gene in the library. To extract the relevant data and generate the plots I wrote the python script `Plot_Log2FC_along_chr.py`. This script requires as input the additional files `EMT5000_library_regions.bed`, and `EMT5000_library_TSS.txt` in order to plot enrichment for each gRNA with respect to the transcriptional start site of the gene from the library. Both files can can be found in the folder `additional_files`. The script can be invoked as follows:

```
python Plot_Log2FC_along_chr.py -c Batch8_p300_DESeq2_table
   .csv -l EMT5000_library_regions.bed -t
   EMT5000_library_TSS.txt
```

This script was used to generate the plots in Figure 4.12 on page 136 of the PhD thesis.

Correlation of Log2Fold enrichment scores from DESeq2 between experiments

The script `plot_LogFC_vs_LogFC.py` performs all-by-all comparison of Log2Fold Change per gRNA across an arbitrary number of inputfiles. The script can be invoked as follows:

```
python plot_LogFC_vs_LogFC.py  [list of DESeq output
   dataframes to be compared]
```

This script was used to generate the plots in Figure 4.13 on page 138 of the PhD thesis.

Identification of candidate gRNAs using ranking with desirability functions

The Log2Fold Change and Log2Fold Change standard error were extracted from the DESeq2 output in python (using ipython notebook interactively) as follows:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import collections as coll
import re
import pylab
import pybedtools
%gui
%matplotlib inline

def load_into_df(arg):
    user_input = pd.DataFrame.from_csv(arg, header=0, sep
       =',', index_col=0)
    return user_input

def set_DESEq2_outliers_to_0_and_wrap_into_df(list_of_dfs,
   names_of_dfs):
    dict_for_df = {}
    for index, df in enumerate(list_of_dfs): #iterate over
       the dataframes
        outliers = df.loc[np.isnan(df['pvalue'])] #make a
           dataframe 'outliers' that holds all rows where
           pvalue is NA, i.e DESeq2 has detected an outlier
        outliers['log2FoldChange']=0 #if pvalue is NA set
           L2FC to 0
        outliers_removed = df.loc[~np.isnan(df['pvalue'])]
           #make new df only containing rows where pvalue
           is NOT NaN
        df_combined = pd.concat([outliers, outliers_removed
           ])   #combine the two dataframes
```

```
        dict_for_df[names_of_dfs[index] + '_log2FoldChange
            '] = (df_combined["log2FoldChange"])
        dict_for_df[names_of_dfs[index] + '_LfcSE'] = (
            df_combined["lfcSE"]) #add the columns holding
            L2FC of this dataframe to a dict
    my_df=pd.DataFrame(dict_for_df) #take entries from a
        list into a df
    return my_df

Batch8_p300 = load_into_df('path2/DESeq2_output_tables/
    Batch8_p300_DESeq2_table.csv')
Batch4_p300 = load_into_df('path2//DESeq2_output_tables/
    Batch4_p300_DESeq2_table.csv')
BatchSS0209_p300 = load_into_df('path2/DESeq2_output_tables
    /BatchSS0209_p300_DESeq2_table.csv')
BatchSS2608_p300 = load_into_df('path2/DESeq2_output_tables
    /BatchSS2608_p300_DESeq2_table.csv')
BatchSS0209_Set7 = load_into_df('path2/DESeq2_output_tables
    /BatchSS0209_Set7_DESeq2_table.csv')
BatchSS2608_Set7 = load_into_df('path2/DESeq2_output_tables
    /BatchSS2608_Set7_DESeq2_table.csv')
Batch5_Set7 = load_into_df('path2/DESeq2_output_tables/
    Batch5_Set7_DESeq2_table.csv')
Batch3_Set7 = load_into_df('path2/DESeq2_output_tables/
    Batch3_Set7_DESeq2_table.csv')

list_p300= [Batch8_p300, Batch4_p300, BatchSS0209_p300,
    BatchSS2608_p300]
names_p300 = ['Batch8_p300', 'Batch4_p300', '
    BatchSS0209_p300', 'BatchSS2608_p300']

list_Set7 = [Batch3_Set7, Batch5_Set7, BatchSS0209_Set7,
    BatchSS2608_Set7]
names_Set7 = ['Batch3_Set7', 'Batch5_Set7', '
    BatchSS0209_Set7', 'BatchSS2608_Set7']

p300s_L2FC = set_DESEq2_outliers_to_0_and_wrap_into_df(
    list_p300, names_p300)
p300s_L2FC.to_csv('p300s_DESeq_L2FC_NApvalue2zero.csv', sep
    =',')
```

```
Set7_L2FC = set_DESEq2_outliers_to_0_and_wrap_into_df(
   list_Set7, names_Set7)
Set7_L2FC.to_csv('Set7s_DESeq_L2FC_NApvalue2zero.csv', sep
   =',')
```

For each gRNA where DESeq2 had detected outliers (and set the p-value to NA in the output) Log2Fold Change values were set to 0 before the next step, i.e. ranking of gRNAs.

The script `Desirability.R` was used to rank gRNAs based on large positive Log2FC and small Log2FC standard error. A weighted average is calculated and Log2FC is given four times the weight of its standard error during this ranking:

```
./Desirability.R -f p300s_DESeq_L2FC_NApvalue2zero.csv -w 4
    -e 1
```

The script outputs a ranked list of candidate gRNAs, ranked based on the calculated overall desirability (here `p300_candidates_4xL2FC_1xLfcSE.csv` as well as plots showing Desirabilities across all gRNAs as well as histograms for Log2Fold Changes and associated standard errors. These are shown in Figure 4.14 on page 139 of the PhD thesis.

The top 10 candidates were extracted from the list of gRNAs ranked by their Desirability score for use in the validation experiments. The sequences were extracted for cloning into the gRNA vector as follows:

```
head -11 p300_candidates.csv | sed 1d | awk -F '",' '{print
    $1}' | awk -F '"' '{print $2}' > p300_top10.bed

grep -A1 -wf p300_top10.bed /path2/
   GN20GG_masked_autoXY_EMT_genepromoter_comprehensive_complete_

noPAM_unique_strand_PAMremoved.fa > p300_top10.fa
```

For the candidate gRNAs the Bayesian-corrected counts in samples and controls were plotted using the custom script `Plot_candidates.R`. For this, the counts for each gRNA for all samples from a given experiment were first extracted from the output of the Bayesian error correction script (see section B.6.2) using the script `extract_all_counts_per_experiment_from_bayes.py`, which was invoked as follows:

```
python extract_all_counts_per_experiment_from_bayes.py
```

This produces, for each experiment a file with extension `_all-counts.csv`. This file was edited to replace all occurrences of NA with 0 using the script `Print_counts_missing_as_0.R`. The file containing the output of the Bayesian error correction script for all experiments using the dCas9-p300 chromatin modifier is shown as an example:

```
./Print_counts_missing_as_0.R -f All_p300_all-counts.csv
```

For this particular example, this script outputs a file named `All_p300_all-counts_missing_as_0.csv`, which contains for all experiments using the dCas9-p300 chromatin modifier the counts per gRNA following Bayesian error correction for all samples and controls with NAs replaced by 0. This file together with the ranked list of candidates (from section B.6.5) was fed to the script `Print_counts_missing_as_0.R` as follows:

```
./Plot_candidates.R -f p300_candidates_4xL2FC_1xLfcSE.csv -
   c All_p300_all-counts_missing_as_0.csv
```

This produced the plots in Figure 4.15 and 4.16 of the PhD thesis.