Techniques and strategies employing engineered transcription factors
Seth Lawsona,b and Leonidas Blerisa,b,c

Abstract
Programmable transcription factors have been instrumental in recent elucidations of developmental and regulatory pathways pertaining to biology and medicine. Programmable DNA-binding domains (DBDs) are cheap and relatively easy to use. Moreover, fusing transcriptional and epigenetic modulators to programmable DBDs can be done with ease. Recent improvements of these modulators have increased the efficacy and appeal of their biomedical applications. Screening and spatiotemporal control of genomic regulation using programmable DBDs have drastically improved the breadth and resolution of (epi)genomic knowledge. The widespread adoption of programmable DBDs for (epi)genomic investigations will lead to a deeper and more comprehensive understanding of diseases and other biological phenomena.

Addresses
a Bioengineering Department, University of Texas at Dallas, Richardson, TX, USA
b Center for Systems Biology, University of Texas at Dallas, Richardson, TX, USA
c Department of Biological Sciences, University of Texas at Dallas, Richardson, TX, USA

Corresponding author: Bleris, Leonidas, Bioengineering Department, University of Texas at Dallas, Richardson, TX, USA. (bleris@utdallas.edu)

Introduction
Irregular epigenetic states are known to underlie human diseases. The ability to perturb these states accurately and efficiently is critical to understanding their mechanisms. The future of personalized and targeted therapeutics will rely on specifically controlling the transcription and chromatin state of DNA for the treatment of these diseases. Current approaches to regulate deleterious epigenetic states rely on non-specific drugs that bear global off-target activity which culminates in cytopathic side-effects. Here we briefly introduce synthetic transcription factors and how they are made, with a focus on Cas9. We review recent advancements in spatiotemporal control and genome wide screening using these tools, and finally conclude with our perspective for the future of synthetic transcription factors and their applications.

The tools
Zinc fingers and TALEs
Cys2His2 zinc fingers (ZFs) are nucleic acid binding motifs consisting of 25 amino acids conferring specificity to overlapping nucleotide quadruplets and a 1–5 amino acid linker [1–3](Figure 1A). ZF arrays can be constructed either by selection-based methods like phage-display, or by recombinant modular assembly, yielding custom DNA-binding domains (DBDs) that can recognize a specific 9–18 nt sequence [4–6]. ZF arrays are fused to transcriptional modulators and then used to perturb programmed target loci [7]. Due to overlapping recognition specificity, ZFs are contextually restricted by neighboring fingers in the array, and many combinations of ZFs are non-functional [2,8,9]. Notably, the Context-Dependent Assembly (CoDA) platform for designing ZFs aids researchers by accounting for neighboring ZFs along the array to choose combinations likely to succeed [10]. For a detailed discussion into the use of ZFs for gene perturbation, Aaron Klug has written a comprehensive review on the subject [11].
(ChIP) of TALE binding sites has revealed sensitivity to local chromatin states [21]. This is ameliorated by using predictive software based on naturally occurring TALE binding sites and considering the chromatin state of the target DNA [16]. Given the complexities and nuances of TALE mediated gene regulation, Moore et al., 2014 provides an extensive discussion on the subject [22].

**Cas9**

The process to design and assemble modular ZF arrays and TALEs is cumbersome, primarily because both require creating a new protein for every desired target (Table 1). Alternatively, the recently elucidated CRISPR-associated protein-9 (Cas9) is a single enzyme that can be targeted to virtually any sequence as determined by an RNA guide (Figure 2A). The Clustered Regularly Interspaced Palindromic Repeats (CRISPR) system is a prokaryotic adaptive immune system [23,24]. The endogenous CRISPR/Cas system destroys foreign phage DNA, and records short fragments of their genomes called protospacers by integrating the sequences into the CRISPR array. These sequences immunize against future infection by guiding nucleases to complimentary targets. Endogenously, the Cas9 system uses a dipartite RNA guide, whereby a CRISPR RNA (crRNA) provides nucleotide specificity and trans-activating crRNA provides a structural handle for Cas9 to associate with [25–27]. Jinek et al. developed a fusion of these two species, a single guide RNA (sgRNA), sufficient for Cas9 activity [27]. The Cas9 endonuclease also requires a pre-determined protospacer adjacent motif (PAM) following the target spacer to initiate cleavage [26–28]. The Streptococcus pyogenes ortholog (SpCas9) has a conveniently short PAM, 5’-NGG-3’, having resulted in widespread adoption and manipulation of this Cas9 ortholog for genome editing. Cas9 will bind to a target that meets the PAM constraint and has complementarity between the guide and the seed region (10–12 nt next to the PAM), though a greater amount of complementarity between the target and guide is necessary for cleavage [29]. Cas9 fused to nuclear import peptides and a sgRNA is all that is needed to target and cleave DNA in mammalian systems [30].

In addition to adapting and reducing the CRISPR/Cas9 system for endonuclease functionality, a nuclease dead Cas9 (dCas9) variant has been generated, with two amino acid substitutions in the RuvC (D10A) and HNH (H840A) nuclease domains [27]. Similar to what was reported by Zhang et al. using TALEs, the spatial accessibility of dCas9’s target also dictates whether a particular sgRNA will function, as well as if an off-target location will be bound or perturbed by (d)Cas9 [29,31,32]. Linker DNA between nucleosome core particles (NCP) is generally accessible to Cas9, whereas the enzyme relies on nucleosomal breathing to bind to targets on or near the NCP [31,32]. Chromatin immunoprecipitation and sequencing (ChIP-seq) has revealed dCas9 binding from tens to thousands of locations with both matching seed sequences and PAM sites [29,33]. However, recent RNA-Seq data using several sgRNAs with dCas9 suggests that these off-target interactions do not yield downstream changes in the transcriptome [34,35]. Therefore, dCas9 functions as a highly versatile DBD, and there is a cornucopia of recently reported dCas9 fusions for modulating gene expression and epigenetic states.

In bacteria, robust repression with dCas9 was obtained by targeting the promoter of a gene of interest (GOI) and sterically hindering the transcriptional machinery, but mammalian genomes were less responsive with an average of ~2 fold repression [36]. CRISPR interference (CRISPRi) was improved with the fusion of the repressor domain Krüppel-associated box (KRAB) [37]. Transcriptional activation using dCas9 (CRISPRa) can be achieved by fusing tandem epitopes of the alpha trans-inducing factor VP16. The greater the number of tandem VP16 domains, the higher the activation. While most studies use VP64 (4 tandem copies), some report as many as 12 tandem copies (VP192) [38]. For stronger activation, multiple dCas9 activators can to the same locus [39–42]. Multiple VP64 domains can be recruited...
to a single locus by linking VP64 domains to single chain variable fragment (scFv) antibodies and appending a peptide scaffold of scFvs to dCas9, a system termed SUperNova (SunTag) [43] (Figure 2B). In an effort to simplify and strengthen Cas9’s ability to target and activate genes, Chavez et al. have created a ternary activation domain by fusing dCas9 to VP64, p65, and Rta (VPR) [44] (Figure 2C). This ensemble is thought to mimic endogenous activation by having multiple activation domains non-competitively recruiting synergistic transcriptional machinery. This is an attractive alternative to localizing several identical activation domains, which intrinsically narrows the scope of transcriptional proteins recruited. Similarly, combining activation domains that recruit a wide variety of factors that modulate transcriptional phenomena, and carries with it applications in cell reprogramming and genetic therapies.

<table>
<thead>
<tr>
<th>DBD</th>
<th>How to program</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>TALE</td>
<td>Clone protein</td>
<td>• Specificity is determined by one repeat per nucleotide</td>
<td>• Must synthesize new protein per desired sequence</td>
<td>2.9–4.4 kb</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Easy to make libraries in a simple molecular biology lab</td>
<td>• Highly repetitive sequences hinder cloning and viral packaging</td>
<td>(12–24 repeats)</td>
</tr>
<tr>
<td>ZF</td>
<td>Clone protein</td>
<td>• Incredibly small size eases AAV packaging and leaves room for other genetic elements to be packaged</td>
<td>• Must synthesize new protein per desired sequence</td>
<td>0.36–0.55 kb</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Widely studied, Quick and inexpensive to program</td>
<td>• Overlap of sequence determination between neighboring ZFs in the array</td>
<td>(4–6 fingers)</td>
</tr>
<tr>
<td>dSpCas9</td>
<td>Synthesize sgRNA</td>
<td>• Can target multiple locations using multiple guides and a single Cas9 protein</td>
<td>• Large protein hinders viral packaging</td>
<td>4.1 kb</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• sgRNA can act as the tether for an effector, permitting different effectors to be directed to different sites within a single cell</td>
<td>• Requires two elements (sgRNA and dCas9) be present to be functional</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Most molecular biology labs must outsource part or all of sgRNA synthesis</td>
<td></td>
</tr>
</tbody>
</table>

Exemplified by sgRNA2.0, the versatility of the sgRNA permits tremendous orthogonality and multiplexing with the Cas9 system [46]. In addition to tolerating aptamer binding protein systems, when sgRNA is truncated to 14–15 nt, catalytically active Cas9 will still assemble with sgRNA and bind to target sites but will not cleave [47,48]. Thus, catalytically active Cas9 can be paired with multiple guides of varying lengths and functionalities to simultaneously induce double stranded breaks and gene regulation, e.g. Cas9 can be combined with a full length guide to cleave one target, while a 14 nt guide with an MS2 loop targets a different sequence and recruits an MCP-VPR fusion using the same Cas9.

**Directed chromatin modulation**

There is a demand for fusions beyond rudimentary effectors like VP64 in order to precisely orchestrate the chromatin landscape. Consider that these domains recruit a wide variety of factors that modulate transcription, imprecisely affecting the chromatin state. Consequently, fusions of DBDs for directed DNA methylation [49–51] and demethylation [50–53], as well as histone methylation [54,55], demethylation [56,57], acetylation [58], and deacetylation [54] have been generated. This suite of epigenome modulators allows researchers to probe location specific epigenetic phenomena, and carries with it applications in cell reprogramming and genetic therapies.

Directing multiple copies of the same epigenetic effector to modulate the same gene (i.e. to the enhancer region and/or promoter of the same gene) has been shown to be ineffective for several fusions [49,52,55,56,58], whereas combining different epigenetic modifiers with similar downstream effects (e.g. a
methylase and a deacetyltransferase) to the same locus results in synergistic effects [55,59].

Medical applications are becoming more feasible as transcriptional and chromatin modulating systems are optimized. Indeed, dCas9-SAM has been shown to efficiently and specifically reactivate latent HIV-1 reservoirs in a variety of CD4⁺T models, which induce host cell death due to cytotoxic viral proteins and eradication by the native immune system [60–62]. This would be preferable to latency reversing agents, which are non-specific and ineffective [63,64]. Potent activators like SAM and VPR have critical implications in treating some of the most common and deadly diseases such as cancer and neurodegenerative conditions [65, 66].

Techniques

Screening

Modular synthetic transcription factors have made the construction of libraries simple and inexpensive, permitting rapid development and testing for user defined loci. Until very recently, biologists have performed library screening assays via RNAi for loss-of-function (LOF) screening and cDNA for gain-of-function (GOF) screening. RNAi suffers from high off-target effects [67,68] and unpredictable or otherwise poor efficiency [69], while cDNA libraries are expensive, often times requiring researchers sacrifice cDNA diversity in lieu of paying to represent multiple isoforms of the same gene [70,71]. With the advent of programmable DBDs, particularly dCas9, library screening has become immensely more viable and efficacious as a strategy for genome-wide perturbation studies (Figure 3A). For guides on constructing and using DBD screens, see Refs. [17,72,73].

CRISPRa/i screens are valuable for studying disease models and various genetic elements in human tissue and animal models [34,45,74–76]. Catalytically active Cas9 has been shown to induce toxicity in eukaryotic cells via double stranded breaks [77,78], as well as elicit host immune response in vivo [79]. Conversely, CRISPRi does not cause non-specific toxicity [80]. Further, despite possibly causing an immune response [81], long-term robust expression of dCas9 and sgRNA has been observed in vivo in the presence of a fully functional immune system [34].

Despite the transient nature of CRISPRa/i, long-term persistent effects are sufficient for many LOF and
GOF applications [34]. Consequently, recent studies employing dCas9-KRAB and dCas9-p300 along with libraries spanning thousands of sgRNAs have successfully screened for gene regulatory elements [82,83]. These groups were able to corroborate reported elements, as well as identify previously unknown regulatory motifs, establishing dCas9-effector screens as powerful instruments to survey the genome. Of medical significance, Konermann et al. used a dCas9-SAM based screen to identify several genes that conferred resistance to the antimelanoma drug PLX-4720 [45]. By transducing a genome-wide library of sgRNAs in dCas9-SAM expressing (BRAF(V600E)) melanoma cells, treatment of PLX-4720 enriched the population for cells which acquired resistance by sgRNA specific gene activation, identifying new and confirming old pathways for PLX-4720 resistance. Even more recently, a dCas9-SAM based genome wide screen found a number of
genes that inhibit influenza A virus (IAV) [84]. A valuable discovery was the B4GALNT2 gene, which strongly inhibited avian IAVs by modifying the α2,3-linked sialic acid-containing glycans on the cell surface, eliminating viral binding. These studies illustrate dCas9’s ability to identify therapeutically significant gene targets in genome-wide screens.

**Spatiotemporal control**

Spatial and temporal control of synthetic transcription factors, as well as catalytically active Cas9, is a critical objective of the scientific community. Future gene therapy solutions will depend on having precise control over the residence time and location of programmable DBDs, while studies of disease and physiology can use this type of control to perturb sensitive minutiae in model systems. Strategies for temporal control focus on limiting the amount of the enzyme present in the cell. An overabundance of Cas9 has shown to increase off-target activity, and it is known that reducing Cas9 induction curtails off-target effects [85,86]. In the case of catalytically active Cas9, self-destructing DNAs can be used to control protein amplitude and duration, thus limiting Cas9 occupancy and decreasing off-targeting [87,88].

Chemical control of synthetic transcription factors is possible when DBDs and transcription factors are divorced and tethered to heterodimerizing proteins [93,94]. TALEs fused to the Rheo Receptor effectively recruit Rheo Activator-VP16 in the presence of GenoStat [95]. Likewise, FK506 binding protein 12 (FKBP12) can be fused to ZFs or TALEs, which can be conditionally dimerized by the FKBP rapamycin binding (FRB) domain fused to transcriptional activators via rapamycin induction [94,96]. Zetsche et al. split the dCas9 protein and fused the moieties to FRB and FKBP12 for rapamycin controlled transcriptional regulation [86] (Figure 3B). The system was shown to be irreversible; a 2 h pulse of rapamycin induced split dCas9-VP64 mediated activation equaling the same as the architecture continuously treated over a 72 h period. Split-Cas9 can also be reconstituted via inteins, whereby the N-terminus and C-terminus fragments are fused to the DnaE C-intein and DnaE N-intein, respectively [81,97] (Figure 3C). When paired, these inteins splice out and ligate their respective proteins to one another, resulting in a full length enzyme [98]. This system permits packaging and reconstitution of the relatively large SpCas9 (4.1 kb) from two separate aden-associated viruses (AAVs), and can be extended to dCas9 [81,97,99]. AAV feature low immunogenicity, and by facilitating the packaging of these tools into AAV, greater functionality and control can be integrated into future DBD-based gene therapies.

**Optogenetics**

Chemical induction gives the user some degree of temporal control. However, chemicals diffuse uniformly and offer no means of spatial control, hampering spatiotemporal genetic studies that require high resolution such as modulation of signaling pathways in specific neurons or disease states sections of tissue. Furthermore, researchers may desire reversible induction. Light-inducible heterodimerization systems offer an attractive alternative; the resolution of spatiotemporal control is on the order of microseconds and micrometers [100], they are non-invasive [101], and they are reversible [102,103]. ZFs and VP16 have been fused to blue light-responsive proteins GIGANTEA (GI) and the light oxygen voltage (LOV) domain of FKFI [104]. TALEs were fused to cryptochrome circadian clock 2 (CRY2) protein, with its heterodimerizing partner cryptochrome-interacting bHLH (CIB1) subsequently fused to various effectors [54] (Figure 3D). These systems respond best to 450–470 nm blue light, and upon induction undergo protein conformational changes that allow binding. Similar to Konermann et al., dCas9 has been used with the CRY2-CIB1 system, with the obvious benefit of its targeting versatility and multiplexability [105,106]. Recently a split dCas9 system, not unlike Zetsche et al., was used in conjunction with photoinducible proteins pMag and nMag [107] (Figure 3E). Since reassembly of dCas9 is made more difficult by fusing large protein domains to the C- and N- termini, pMag and nMag have the advantage of being smaller (both 150 amino acids) than CRY2:CIB1, presumably permitting reconstitution of the full length enzyme with less steric hindrance. Thus far, the appeal of optogenetic control has been to specifically perturb single pathways in a light controlled manner, yet it is conceivable that photoinducible control of dCas9 will be employed with multiplexed sgRNA circuits to initiate orthogonal transcriptional changes throughout the genomes of the model in question, with applications in neuroscience and 3 dimensional disease models.

**Conclusions**

The application of genome-scale libraries and multiplexing using synthetic transcription factors have been complemented by falling costs of high-throughput sequencing [108]. This synergy is permitting researchers to unravel the epigenetic states underlying medical conditions and drug interactions in a novel and high-throughput manner. Given that such multiplexed screens will identify gene therapy solutions that call for orthogonal up- and down-regulation of multiple targets, DBD-effectors require several improvements to become a
viable therapy. Smaller Cas9 variants are needed if such hypothetical multiplexed programs are to be packaged as Cas9-AAVs, which investigations to date have sought through computational discovery [109] or synthesized by rational design i.e. the removal of the “expendable” REC2 domain of SpCas9 [110]. Indeed, the number of available molecular tools for genomic and epigenomic exploration is increasing rapidly. New CRISPR proteins with unique PAMs and functionalities are still being discovered such as the recently described Cpf1, which shows promise as a DBD for coupling to effectors [111–113]. Difficulty lies in achieving comparable activity to SpCas9, as many orthologues are less efficient or don’t work at all in eukaryotic systems, and the REC2-negative truncated Cas9 exhibited 50% activity of wild-type. Like we have observed with the generation of new PAM variants [114], selection based methods are a potential solution.

Beyond packaging constraints, off-target effects and context dependent efficiencies are of concern for all DBDs and catalytically active Cas9 when considering their viability as a potential therapy. Despite some studies little to no off-target effects to the transcriptome when using DBDs [34,35,115,116], adoption of dCas9 and other DBDs for gene therapies will require further study. Techniques like using paired nickase mutants greatly reduces off-target effects for catalytically active Cas9, but easier to use options like the recently engineered Cas9s with greater specificity [117,118] have yet to be tested extensively in vivo. Lastly, resolving the chromatin state sensitivity of these proteins would be a helpful in making single target and pooled assays more consistent and convenient to users, since many targets aren’t able to be bound by these DBDs due to local chromatin structure.

Engineered transcription factors have expanded our ability to probe and control the (epi)genome with unprecedented precision and versatility. These tools are efficacious for most basic biology studies in their current form; however, the problems outlined here are critical if engineered transcription factors are to be universally adopted and serve as the next generation of gene therapies.

**Acknowledgements**

This work was funded by the US National Science Foundation (NSF) CAREER grant 1351354, NSF 1361355, and the University of Texas at Dallas. We thank Chance Nowak and Ian Kampine for assistance in editing this manuscript.

**References**

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest


35. Authors conduct a dCas9 screen to identify previously unknown genomic sites that dictate tumor growth and drug response. Interestingly, they demonstrate dCas9-VP64 can either activate or repress transcription depending on its target site relative to the transcription start site.


37. VP64 and second generation activators are compared, in which SAM is found to be the superior. Attempts to combine SunTag, VPR, and SAM are unsuccessful, and other data suggest an inherent upper limit of each system to activate a gene.


48. Authors design sgRNA2.0, where RNA aptamers are engineered into the sgRNA. In addition to developing SAM, sgRNA2.0 can recruit virtually any protein or combinations of proteins to a target entirely determined by the sgRNA.


A contemporary example of using CRISPRi screening, where they elucidated the role of IncRNA in cell growth and cancer.


Avery comprehensive guide to making and using (d)Cas9 screens.


Algorithm considers various contexts of dCas9 target sites to generate an efficient library of sgRNAs for human and mouse editing.


Recent example of identifying regulatory elements using dCas9 based high-throughput screening. The authors use both activation (p300) and repression (K/RAB) screens to cross validate results.


Authors used a massive sgRNA library spanning 1.29 Mb of DNA, demonstrating a brute force method for identifying regulatory elements.


