

Iterative strategy for discovering *in vitro* directed cell state transitions using combinatorial activation and repression of endogenous genomic loci

Gleb Kuznetsov — Biophysics 205 Final Project Proposal — April 28, 2015

Abstract

Improved strategies for cellular reprogramming and directed differentiation offer to advance the potential for stem cell-based therapies and patient/tissue-specific drug screening. Many examples of induced cellular state transitions have been reported, with derived cell types resembling their target cell types morphologically and functionally. However, reprogramming is incomplete in most cases, and common methods such as ectopic factor expression and shRNA knockdowns are fundamentally low throughput. Notably, most induced cell lines have been shown to harbor expression signatures of their originating cells [1], which may ultimately limit the stability of their fate and therapeutic potential. I propose leveraging advances in multiplexed control over cell epigenetic state in order to combinatorially explore reprogramming and directed differentiation pathways with the goal of identifying cell fate modification programs that are less constrained by assumptions and knowledge *a priori*, and yield cell types that are closer to their target types on a molecular level. I describe an iterative strategy for discovering such pathways. This strategy leverages catalytically dead Cas9 (dCas9) fused to activating or repressing domains along with a combinatorial library of guide RNA arrays simultaneously targeting activation and/or repression of multiple genomic loci. Fluorescence-activated cell sorting (FACS) provides a high-throughput means of assaying progress towards a goal followed by higher resolution inspection of cell state through transcriptional profiling to bring engineered cells closer to their target cell types.

Introduction

Better strategies for deriving cell types of interest *in vitro* will play a central role in regenerative therapies and patient- and tissue-specific drug screening. The field of reprogramming gained much attention with the first successful cloning of a mammal nearly two decades ago [2]. A decade later, it was shown that somatic cells can be induced into an embryonic-like state through the ectopic expression of just four factors: Oct4, Sox2, Klf4, and c-Myc (OSKM) [3], now commonly referred to as the Yamanaka factors. Today there exists a multitude of examples of somatic cells being reprogrammed to pluripotent stem cell state and other examples of directed differentiation to various progenitor and terminate cell fates.

Still, while reported reprogrammed cells resemble their target types on a morphological, functional, and molecular level, limitations in both the methods of deriving these cells and the shortcomings of reprogrammed cell products in faithfully reproducing their targets remain obstacles in the pursuit of therapies and drug screening efforts. A commonly employed technique for reprogramming, the overexpression of ectopic factors, is limited in throughput and results in the integration of exogenous genetic material into the genome of a cell, limiting the clinical utility of the engineered cells and potentially confounding the results of drug screening studies. Methods that avoid viral integration through the use of mRNAs [4], are technically more complex and lower throughput, limiting their potential as a platform for discovering improvements to reprogramming pathways.

A pair of complementary studies recently demonstrated quantitatively how most engineered cell lines examined still carry significant gene expression signatures of the cell lines they originated from [1, 5]. The software platform at the core of these studies, CellNet, infers the gene regulatory networks (GRNs) active in a cell, currently a gold standard for quantitative representation of cell identity [6], and reports deviations from target tissue type. Despite painting a less than ideal picture of the state of the reprogramming field, CellNet is capable of quantitatively informing follow-up engineering efforts by recommending potential target factors for up- or down-regulation that may yield higher fidelity reprogrammed cell lines. An iterative process of ectopic expression or knock down of targets recommended by CellNet was demonstrated as a fruitful method for dissecting cell fate in applications such as improving B cell to macrophage conversion and clarifying that cells previously reported to be induced hepatocytes as actually being endoderm progenitors [5]. However, recent advances in technologies to perturb cell epigenetic state potentially offer a higher throughput means of leveraging data on GRN differences between engineered and target cells.

The CRISPR-Cas9 system is an RNA-programmable endonuclease, whose DNA-cutting activity can be inactivated to yield a precise genomic homing device. Such a catalytically dead Cas9 (dCas9) can be fused to effector domains such as activators, repressors, or chromatin modifiers yielding a tool for precise epigenetic control [7]. Several recent advances have improved the activating strength of dCas9-activating domain fusions [8,9], yielding a tool that provides up to four magnitudes of activation of endogenous loci over baseline expression levels. Additionally, the CRISPR-Cas9 system is naturally multiplexable through the specification of multiple guide RNAs (gRNAs) to direct the protein to multiple genomic targets in parallel. Finally orthogonal Cas9 proteins [10] can be fused to different effector domains, allowing simultaneous execution of functions of distinct nature (e.g. simultaneous activation and repression) at distinct loci.

The ability to limit the space of combinations to test through a quantitative understanding of cell GRNs enabled by CellNet and the availability of highly programmable dCas9-effector domain fusions presents an opportunity for an iterative, high-throughput method for improving cell reprogramming and directed differentiation. I describe the development and application of this method in a series of three aims of increasing complexity and discovery potential: In *Aim 1*, I focus only on testing combinations of TF activation and validate through an experiment to “rediscover” the Yamanaka factors in the induction of mouse embryo fibroblast to pluripotent embryonic stem cell-like fate [3]. In *Aim 2*, I extend the method to allow

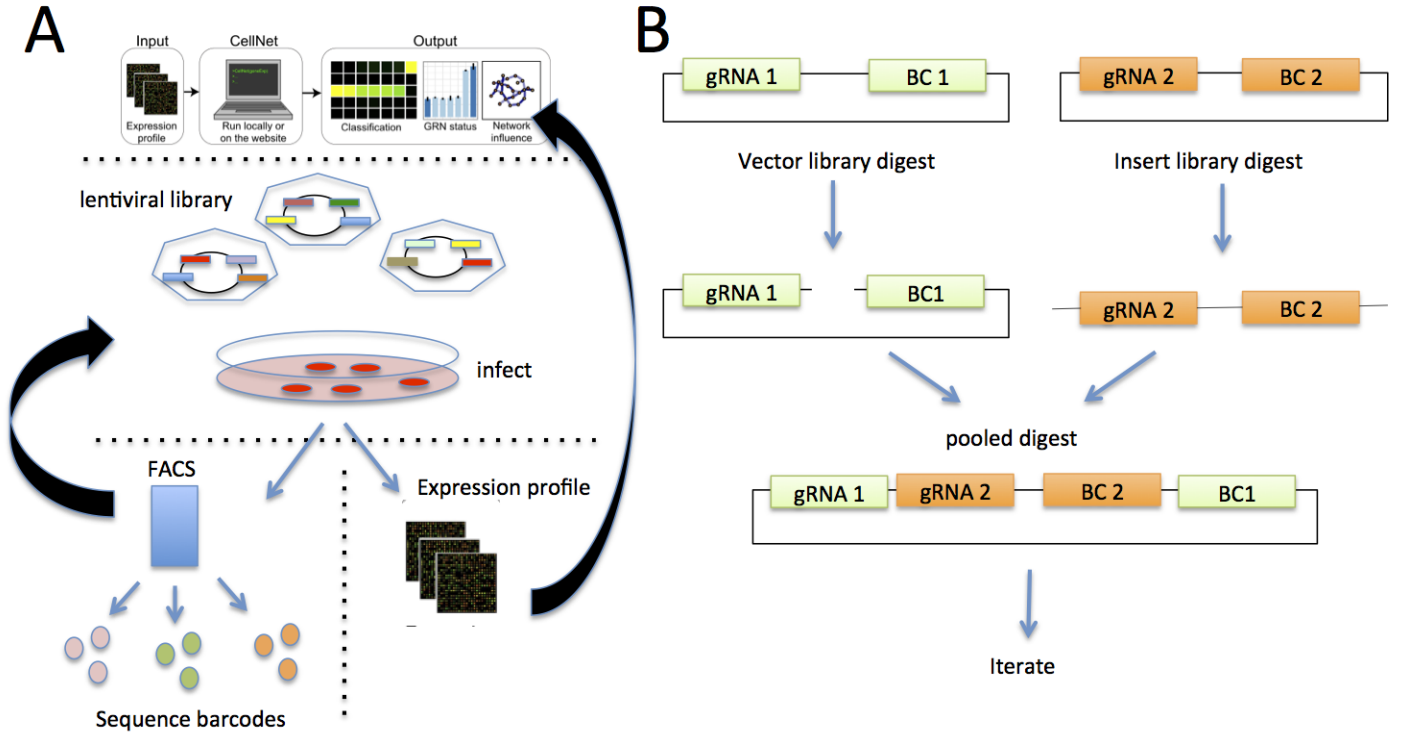


Figure 1: (A) Overview of iterative reprogramming approach using lentiviral library composed of combinatorial gRNAs. (B) Illustration of combinatorial cloning strategy based on CombiGEM [11] to construct libraries of barcoded gRNA arrays.

simultaneously searching over activation and repression combinations by leveraging orthogonal dCas9 [10] separately fused to activation and repression domains. I validate this portion of the system in the context of B cell to macrophage conversion. Finally, in *Aim 3*, I describe an open-ended method of hierarchically discovering de novo reprogramming pathways through searching a large combinatorial space of activation and repression combinations and assaying for the appearance of cell state markers of increasingly terminal fate in successive iterations of experimentation.

Experiments are carried out in mouse to leverage the datasets already available in CellNet [1] and to be able to reproduce previously reported reprogramming and directed differentiation results in developing this method.

1 Aim 1: Test combinatorial gene activation only and validate by “rediscovering” Yamanaka factors.

The purpose of this aim is to test the essence of the iterative method described in this proposal, by limiting attention to combinatorial gene activation only. To validate and tune the method, I will carry out the exercise of “rediscovering” the Yamanaka reprogramming factors Oct4, Sox2, Klf4, and c-Myc (OSKM) that were shown to induce mouse embryonic fibroblasts into a pluripotent state [3].

Figure 1 shows a general overview of the approach consisting of the following steps:

1. Identify the candidate transcription factors for which we’ll test combinations. For *Aim 1*, I start with the 24 candidate factors used by Takahashi and Yamanaka [3].
2. Prepare lentiviral library to deliver dCas9-effector fusion and combinatorially assembled gRNA array.
3. Infect cells and incubate.
4. Harvest cells and sort using FACS, targeting markers important in target cell type.
5. Sequence barcodes for each well and use enrichment information to iterate on construction of new, more focused lentiviral library. Repeat steps 3-5 until no further significant enrichment is observed.
6. Perform transcriptional profiling and analysis with CellNet to interrogate GRN state. Use this information to update lentivirus library and continue to iterate.

1.1 General Design Decisions

There are a few design decisions to make *a priori* but these will likely need to be tuned throughout developing the method. One decision is which dCas9-activator fusion to use. Two powerful systems were recently reported: an engineered dCas9-activator system designed by inspecting dCas9 structure from the Zhang Lab [8] and a dCas9 fused to a tripartite activator, VP64-p65-rta (VPr) from the Church Lab [9]. Both were shown to have as much as four orders of magnitude-fold activation over baseline expression for some targets. An alternative option is to use a dCas9-acetyltransferase fusion that acetylates histone H3K27, allowing activation at enhancers [12].

Another design decision to make is how many gRNAs to test in parallel. This decision is constrained by the demonstrated decreasing activation effect with increasing target multiplexity [8,9]. Another constraint is imposed by the Combinatorial Genetics En Masse (CombiGEM) method [11] (detailed below) used for constructing the gRNA-barcode array element. Additionally, it's necessary to think about the combinatorial library size. Guided by the *Aim 1* goal of rediscovering the four Yamanaka factors and the highest multiplicity reported for CombiGEM so far of barcoded elements, I will start with cassettes consisting of four guide RNAs.

Finally, another design decision is what multiplicity of infection (MOI) to use for lentiviral infection. For simplicity, I will start with a MOI of one, although increasing this number increases the opportunities for testing combinations of higher numbers of combinations, if not exhaustively.

1.2 Controls

Given knowledge of the OSKM “solution” to the search for the Yamanaka factors, a positive control will be carrying out the steps of the experiment with the individual gRNA component inputs to the combinatorial library construction limited to those targeting only the four OSKM factors. The negative control will be the complementary library that omits gRNAs targeting any of the OSKM factors. One caveat, discussed below under *Risks*, is that it's not clear that a single guide RNA (sgRNA) targeted to the promoter region of any of OSKM is sufficient for adequate activation. To verify adequate expression for the best-predicted sgRNAs for OSKM, I would conduct quantitative PCR on cells transduced with the OSKM and compare to non-transduced cells to verify activation of the target loci.

1.3 Steps in Depth

1.3.1 Identifying candidate transcription factors for screen

Since the number of transcription factors in the human genome numbers around 2600 [13], the space of possible combinations becomes intractably large, even at a combination multiplicity of two. Instead, it is necessary is to computationally assess expression data for the target tissue type and/or leverage CellNet [1] to generate a reduced list of candidate nodes that may play a role in establishing the gene regulatory networks for the tissue type.

However, given the main purpose of *Aim 1* being to serve as a proof of concept of the method end-to-end, I will artificially narrow the space of candidate transcription factor targets to the 24 candidates originally targeted by Takahashi and Yamanaka [3]. To identify and design gRNAs, I will use guidelines and computational resources provided in [8] to choose the best candidates. If a single gRNA is used for each target, and we use multiplicity of 4, this yields a library size $\binom{24}{4} = 10626$ combinations. However, a caveat to point out is that it may be necessary to use multiple guide RNAs for a particular target, which may bring challenges into library size considerations.

1.3.2 Constructing combinatorial lentiviral library

Genetic constructs will be delivered to cells by lentivirus transduction. The choice of lentivirus for delivery is guided by the opportunity to have reasonable control over multiplicity of infection through modulating virus titer vs cell concentration, as well as stable integration of the payload, enabling propagation throughout cell progeny. Ultimately, it will be desirable to introduce the discovered perturbations without viral integration [4].

The payload delivered by a given lentivirus includes a an array of gRNAs, an array of corresponding barcodes that uniquely identify the combination and order of gRNAs, the dCas9-effect fusion, and a selectable marker for being able to purify for cells that are successfully infected.

To construct the combinatorial library of gRNA and barcodes, I will use the strategy of Combinatorial Genetics En Masse (CombiGEM) [11] originally developed for use in *E. coli* and recently communicated to function in mammalian cells [14] (Fig. 1B). Briefly, CombiGEM uses an iterative cloning strategy starting with a library of DNA elements consisting of a gRNA and corresponding barcode. The DNA element includes restriction sites between the gRNA and barcode and flanking each. A library of *vectors* is generated through cleaving at the sites between gRNA and barcodes, and a library of *inserts* is created by cleaving around the gRNA and barcode. The vector and insert libraries are combined and ligated, yielding new DNA elements with an array of gRNAs and the corresponding array of barcodes. This process is repeated for each order of multiplexity desired, and was reported to work robustly to at least a multiplexity of four in [11].

1.3.3 Infection with lentivirus and incubation

The lentiviral library is delivered to the cells at the desired multiplicity of infection (MOI). The simplest strategy would be to use an MOI of one; however, increasing the MOI potentially allows testing greater combinations. After some time (48 hours for Yamanaka context), cells are challenged by compound for which selectable marker delivered by lentivirus protects against. Cells are allowed to incubate for more time (48 hours for Yamanaka context), allowing for propagation of expression changes and ideally appearance of reprogrammed cells.

1.3.4 Harvesting cells and FACS

Knowledge of the target GRNs informs selection of antibodies bound to fluorophores for FACS. For the Yamanaka factor rediscovery experiment, there are many commercially available kits for assaying induced pluripotent stem cells. These include target markers of pluripotent state such as OCT4, SOX2, SSEA4, and TRA-1-60. The appropriate markers should be determined based on the cell type of interest. It is possible to use more than just two channels in order to bucket by expression at varying intensity.

1.3.5 Sequencing barcodes, designing next lentiviral library

Following FACS, the barcode regions are amplified using pre-designed PCR primer adapters [11] and the barcodes are read using next generation sequencing. The resulting data is analyzed for enrichment of targeted factors. Based on enrichment, the lentiviral library is redesigned to over-represent the enriched factors. Repeat steps 3 - 5 several times until no significant change in enrichment is observed.

1.3.6 Transcriptional profiling, inferring GRNs

After the final iteration of lentiviral infection, the best expressing cells as assessed by FACS can be transcriptionally profiled to gain a higher resolution understanding of cell state. Bulk chip or RNA-seq may be adequate given the sorting carried out, but single cell RNA-seq may yield insights that bulk sequencing does not, especially allowing connecting expression to specific barcodes. Comparing the GRN of these final cells to the target GRN state will inform additional factors to test, which may have been left out of the original screen. A new lentiviral library can then be redesigned and infection and sorting and/or transcriptional profiling repeated.

1.4 Risks

Activation efficiency: dCas9-activating domain fusions have only been characterized for a handful of factors and already this has shown that the amount of activation depends on the gene as well as the specific gRNA target [8,9]. Additionally, activation fold falls with increasing multiplexity. Nonetheless, it has been shown that a single guide RNA is sufficient for significant activation of several factors [8].

Calibrating lentivirus multiplicity of infection: The current bulk barcode sequencing strategy does not give a way to test multiplicity of infection.

Biological limitations: It might be possible that certain loci are inaccessible due to chromatin state and none of the starting pool of transcription factors can serve as an adequate pioneer factor to initiate transition.

2 Aim 2: Implement searching over combinations of activation and repression

In this aim, I add multimodality to the combinatorial method developed in this proposal, introducing the ability to simultaneously activate and repress different loci. To validate this method, I use B cell to macrophage conversion, a relatively quick (10 hours) and robust conversion that has been reported to only require a single transcription factor C/EBP α [15]. However, CellNet analysis revealed that a higher fidelity conversion analysis additionally required the knockdown of Pou2af1 and Ebf1 [5]. The requirement of simultaneous up- and down-regulation, small number of factors, and rapid conversion reduces the variables in testing a multimodal combinatorial strategy.

To allow simultaneous activation and repression of different loci, I will create orthogonal dCas9s based on orthogonal Cas9 proteins reported in [10]. Orthogonal Cas9 proteins, which recognize different RNA scaffolding structures and different PAM sequences, allow simultaneously carrying out different functions on distinct targets in the genome. The corresponding gRNA library designs must use the correct scaffolding RNA elements (crRNA and tracrRNAs) for their respective dCas9.

CellNet will be used to specify which factors in the original cell type should be down-regulated and which should be up-regulated. For validating *Aim 2*, this should include the solution factors found, but also include additional factors in order to show that a solution can be found. Two separate lentiviral libraries will be created, one for activation and the other for repression. Infection can be carried out in parallel. Despite looking for combinations of both activation and repression,

it's not necessary that all cells get both as even just infection with an activating library may bring a cell type closer to the target.

A caveat that arises here is that the current strategy doesn't allow for identifying which pairs of up-regulation and down-regulation libraries co-occur. Still, limiting the library based on iteration as in *Aim 1* should help clarify this.

2.1 Controls

Similar to the Yamanaka system above, as a negative control I would leave out the solution elements C/EBP α as activator targets and Pou2af1 and Ebf1 as repressor targets and look for underrepresentation of macrophage factors in FACS. As a positive control, I will construct libraries enriched the previously left-out factors and expect higher proportion of converted cells.

2.2 Risks

How orthogonal is orthogonal?: Are the Cas9s truly orthogonal with respect to binding their target gRNAs? **Multiple infection by lentivirus:** Are there issues that arise from multiple lentivirus infection?

3 Aim 3: Hierarchical discovery of do novo reprogramming pathways

This aim is intentionally open-ended and involves pushing the limits of the iterative method developed in this proposal. The goal is to identify novel pathways for directed differentiation by assaying for increasingly differentiated cell types in subsequent rounds of lentiviral library design, infection, and FACS.

The process start with a set with mouse embryonic stem cells and a large library of gRNAs. The same locus may be targeted for activation and repression in the orthogonal lentiviral libraries. Activating gRNAs should be enriched for pioneer factors, as well as factors that serve as hubs in more GRNs. Cycles of lentiviral library design, infection, and FACS are carried out as before, with the additional detail of updating the set of antibodies using FACS to select for increasingly differentiated cell types. In the first run of FACS, a broad set of antibodies are chosen, specific to general lineage specification marks, e.g. unique markers for endoderm, ectoderm and mesoderm. Following FACS, the respective barcodes in each bucket should be enriched for their respective fate specification. Three new pools of lentiviral factors are created, each respectively enriched for markers in the respective endoderm, ectoderm, and mesoderm buckets. The process is then repeated for the three separate pools. For each pool, a different set of the next-most differentiated antibody types is selected. The branching factor of this process starts to get large so only the most interesting lineages can be followed, to be chosen based on the most promising hits in subsequent rounds.

3.1 Controls

As a positive control, use limited libraries targetting endoderm, ectoderm, and mesoderm lineage-specific factors to ensure enrichment of these types. As a negative control, target only a single lineage factor, to ensure that other lineages are not spurious hits.

3.2 Risks

Large library size: The size of the initial gRNA (TBD) may be too large to find a tractable solution. Still, individual factors may be enough to push down interesting paths.

General Caveats

Role of Cell Environment: The environment of cells is known to play an important role in cell fate [16]. The experiment setup in this proposal does not take this into account at the moment. It may be possible that obtaining a particular cell type requires the surrounding environment, including cells and physical setup, to be of a certain distribution of cells.

Relative expression As designed, the strategy in the current proposal has not taken into account how to tune relative expression of factors. The effect of dCas9-effector fusions depends on gene context and choice of gRNA, among other factors [8,9]. The relative expression of each factor, especially during transition, likely matters and may not be evidenced by the final relative expression observed in the target cell type.

Role of temporal expression: The strategy in this proposal does not take into account the temporal dimension of transcription factor expression. One step in the right direction is to recapitulate development by using multiple lentiviral libraries that contain orthogonal dCas9-effector fusions that are also orthogonally inducible, allowing sequential expression.

References

1. Patrick Cahan, Hu Li, Samantha A Morris, Edroaldo Lummertz da Rocha, George Q Daley, and James J Collins. Cellnet: network biology applied to stem cell engineering. *Cell*, 158(4):903–915, 2014.
2. Keith H Campbell, Jim McWhir, WA Ritchie, and Ian Wilmut. Sheep cloned by nuclear transfer from a cultured cell line. *Nature*, (380):64–6, 1996.
3. Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126(4):663–676, 2006.
4. Luigi Warren, Philip D Manos, Tim Ahfeldt, Yui-Han Loh, Hu Li, Frank Lau, Wataru Ebina, Pankaj K Mandal, Zachary D Smith, Alexander Meissner, et al. Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mrna. *Cell stem cell*, 7(5):618–630, 2010.
5. Samantha A Morris, Patrick Cahan, Hu Li, Anna M Zhao, Adrianna K San Roman, Ramesh A Shivdasani, James J Collins, and George Q Daley. Dissecting engineered cell types and enhancing cell fate conversion via cellnet. *Cell*, 158(4):889–902, 2014.
6. Eric H Davidson and Douglas H Erwin. Gene regulatory networks and the evolution of animal body plans. *Science*, 311(5762):796–800, 2006.
7. Lei S Qi, Matthew H Larson, Luke A Gilbert, Jennifer A Doudna, Jonathan S Weissman, Adam P Arkin, and Wendell A Lim. Repurposing crispr as an rna-guided platform for sequence-specific control of gene expression. *Cell*, 152(5):1173–1183, 2013.
8. Silvana Konermann, Mark D Brigham, Alexandro E Trevino, Julia Joung, Omar O Abudayyeh, Clea Barcena, Patrick D Hsu, Naomi Habib, Jonathan S Gootenberg, Hiroshi Nishimasu, et al. Genome-scale transcriptional activation by an engineered crispr-cas9 complex. *Nature*, 2014.
9. Alejandro Chavez, Jonathan Scheiman, Suhani Vora, Benjamin Pruitt, Marcelle Tuttle, Eswar Iyer, Shuailiang Lin, Samira Kiani, Christopher Guzman, Daniel Wiegand, and et al. Highly efficient cas9-mediated transcriptional programming. *Nature Methods*, 0, 2015.
10. Kevin M Esvelt, Prashant Mali, Jonathan L Braff, Mark Moosburner, Stephanie J Yaung, and George M Church. Orthogonal cas9 proteins for rna-guided gene regulation and editing. *Nature methods*, 10(11):1116–1121, 2013.
11. Allen A Cheng, Huiming Ding, and Timothy K Lu. Enhanced killing of antibiotic-resistant bacteria enabled by massively parallel combinatorial genetics. *Proceedings of the National Academy of Sciences*, 111(34):12462–12467, 2014.
12. Isaac B Hilton, Anthony M D’Ippolito, Christopher M Vockley, Pratiksha I Thakore, Gregory E Crawford, Timothy E Reddy, and Charles A Gersbach. Epigenome editing by a crispr-cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature biotechnology*, 2015.
13. M Madan Babu, Nicholas M Luscombe, L Aravind, Mark Gerstein, and Sarah A Teichmann. Structure and evolution of transcriptional regulatory networks. *Current opinion in structural biology*, 14(3):283–291, 2004.
14. Timothy K Lu. submitted, reported at mammalian synbio conference april 26, 2015. 2015.
15. Lars H Bussmann, Alexis Schubert, Thien Phong Vu Manh, Luisa De Andres, Sabrina C Desbordes, Maribel Parra, Timo Zimmermann, Francesca Rapino, Javier Rodriguez-Ubrea, Esteban Ballestar, et al. A robust and highly efficient immune cell reprogramming system. *Cell stem cell*, 5(5):554–566, 2009.
16. Daniela Drummond-Barbosa. Stem cells, their niches and the systemic environment: an aging network. *Genetics*, 180(4):1787–1797, 2008.

Statement of Outside Help

Daniel Goodman (PhD student, Church Lab) - Conversation regarding synthesizing combinatorial gRNA cassette libraries. Discussed limitations on Chip-based synthesis, as well as strategies (barcodes, etc.) for read-out.

Alex Ng (PhD student, Church Lab) - Conversation regarding issues related to combinatorial targeting using dCas9-activating domain fusion. Alex pointed out the caveat that activation strength for any given promoter is some amount relative to the baseline of that promoter.