ARCHIVES
MASSACHUSETTS INSTITUTE OF TECHNOLOLGY
JUN 12 2015
LIBRARIES

GENOMIC NUCLEIC ACID MEMORY STORAGE WITH DIRECTED ENDONUCLEASES

BY
NOAH JAKIMO

Bachelor of Science in Computer Science, California Institute of Technology, 2010

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

© Massachusetts Institute of Technology 2015. All rights reserved.

AUTHOR Signature redacted .....
Program in Media Arts and Sciences
May 8, 2015

CERTIFIED BY Signature redacted .....
Joseph Jacobson
Associate Professor of Media Arts and Sciences
Thesis Supervisor

ACCEPTED BY Signature redacted .....
Pattie Maes
Academic Head
Program in Media Arts and Sciences

GENOMIC NUCLEIC ACID MEMORY STORAGE WITH DIRECTED ENDONUCLEASES

BY NOAH JAKIMO

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning on May 8, 2015 in partial fulfillment of the requirements for the degree of Master of Science.

Abstract

Technologies for long-term recording of cellular pathway activation are constrained by the difficultly to constantly monitor transient signaling events and expression of target genes. To overcome these limitations we designed a recording tool that uses the transcriptional output of a signaling pathway as the input for an engineered genome-encoded memory. The mechanism of recording leverages the programmable nature of the bacterial immune system that consists of Clustered Regularly Interspaced Short Palindromic Repeat Sequences (CRISPR), which can recognize and cleave viral DNA using an RNA-guided directed endonuclease. Cuts left by the endonuclease are repaired by an error-prone DNA damage repair mechanism, namely non-homologous end joining (NHEJ), likely to leave mutations at the cut sites. Defining the cut site with pathway-dependent transcription of guide RNA, this genomic region is sequenced to measure pathway activation by the amount of accumulated mutations. To demonstrate a system to monitor cancer metabolism, guide RNA is expressed in mammalian cell culture with a NF-kappaB promoter. To demonstrate a system that can monitor sugar intake in an environment like the gut, guide RNA is expressed in bacteria with an arabinose promoter.

Thesis Supervisor: Joseph Jacobson
Title: Associate Professor, Program in Media Arts and Sciences

GENOMIC NUCLEIC ACID MEMORY STORAGE WITH DIRECTED ENDONUCLEASES

BY NOAH JAKIMO

The following people served as readers for this thesis:

Signature redacted

Neil Gershenfeld
Director, MIT Center for Bits and Atoms

Signature redacted

Ed Boyden
Associate Professor of Media Arts and Sciences, MIT

Signature redacted

David Sabitini
Professor of Biology, MIT

Acknowledgements

I would like to thank the following groups of people (listed in alphabetical order within group) for their important contributions to the work described in this thesis:

Collaborators:

Dr Naama Kanarek, Lisa Nip

Supervisors:

Prof Joseph Jacobson, Prof David Sabatini

Contributors:

Dr Shmulik Motola, Dr Omri Wurztel

Reading Committee Members:

Prof Ed Boyden, Prof Neil Gershenfeld

Table of Contents

Chapter 1:
Introduction

"The Central Dogma: This states that once 'information' has passed into protein it cannot get out again. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible."

-Francis Crick, 1958

Next-Generation Central Dogma of Molecular Biology

Nucleic acid and peptide chains are all forms of digital information storage. The central dogma of molecular biology describes how such digital information flows between deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins. By extending the central dogma to data storage in electromagnetic devices, bits can written to DNA by chemical synthesis and read from DNA through sequencing [Gillings, 2014] .

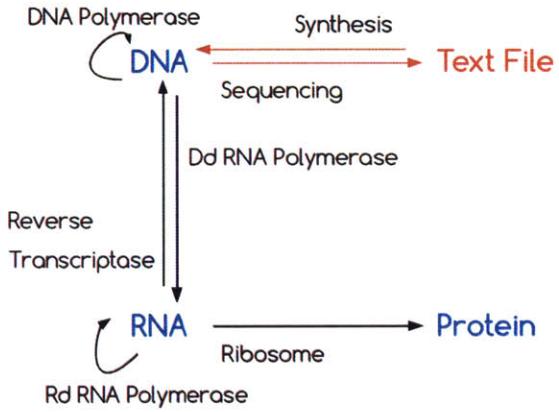


Figure 1. The central dogma of molecular biology, which covers DNA, RNA, and protein, can be extended to computer-based storage of nucleic acid sequences, such as text files. Adapted from Gillings, 2014.

The current global capacity for DNA sequencing can be repurposed exclusively to sequence the genome of every living person nearly every two hours [Carlson, 2014]. Enough bases are also synthesized per person daily to compose a whole genome of a novel bacteria an individual may like to inject into their microbiome or environment. Therefore, especially in light of fields like biomedical research and synthetic biology, the electronic-DNA interface continues to become increasingly important and accessible.

While manipulating nucleic acid sequences on a computer is suited for genomic analysis and design, changes in nucleic acid sequences on DNA in response to biochemical pathways make DNA suitable for recording signals within a biological cell. This thesis combines the advantages of both mediums to engineer a system that stores information from the cellular environment into DNA, propagates the information through cell lineages by replication, and can be later measured through DNA sequencing.

Reasons to Store Memory in Replicating DNA

Why store information locally in a living cell's DNA rather than recording traditional measurements of the cell by microscopy, spectroscopy, electrophysiology, etc. onto hard drives? To answer this question, first consider the cost of manufacturing a light -detecting photodiode versus growing a cell that expresses photosensitive proteins, like channelrhodopsins.

Figure 2. Left: Silicon and Germanium photodiodes ; Right: E. Coli viewed at 10,000x from an electron microscope. Both images are re-used via Creative Commons.

The catalog of Digi-key, an electronics distributor, lists phototransistors for as little cost as $0.05875 – when purchased in bulk orders of at least 3,000 units. Suppose then that the manufacturing cost is no more than $0.05 [DigiKey, 2014].

For editing recombinant DNA into a plasmid, *E. coli* bacteria cultured at 37C in growth media for several hours can efficiently generate many copies of the plasmid. A datasheet of a compact incubator supplied by Thermo Scientific, a biological equipment manufacturer, lists a 14W power consumption for maintaining 37C [Thermo, 2010]. Given an electricity cost of $0.17 / kWh in Boston according to the U.S. Bureau of Labor Statistics and a necessary incubation time of 12 hours before the completion of log phase growth in the bacterial culture, together suggests an electric bill of $0.02856 [USLabor, 2014]. New England BioLabs, a biochemical supplier, sells 100 mL of SOC Outgrowth Medium for $71.00 [NEB, 2014]. Since cells are harvested at a density of 4e9 cells / mL, it then follows that the cost of growing each cell is roughly $1.75e-10.

On one hand, growing bacteria is by far cheaper than fabricating photodiodes. On the other, according to the photodiode's datasheet and literature in optogentics and bacterial growth rates, the photosensitivity and temperature range of the photodiode are roughly twice as broad as those of *E. coli* co-expressing multiple channelrhodopsins. Hence the decision to fabricate or grow a sensor is strongly tied to operating conditions and the range of the desired signal.

Now consider several other properties of *E. coli*: 1) they move at speeds of 116 um/s; 2) their genome consists of 4.2e6 basepairs; 3) they replicate once in about every 30 minutes with a mutation rate of one in 10e10 nucleotides. In other words, constant monitoring under a light microscope equipped with a camera is necessary to simultaneously track each *E. coli* cell within a growing population [Phillips, 2012]. Even so, the population will expand beyond the field of view after a sufficient number of hours. By coupling the channelrhodopsins' absorption of photons to certain mechanisms that modify DNA, information on each cell's light exposure can be stored

locally into the genome rather than having to perpetually image every organism at once.

Local storage in DNA is an especially critical feature for enabling single-cell analysis in many applications. One example application is covered in a review last year by Marblestone et al, who analyzed of a multitude of approaches to measuring concurrent electrical activity in each of the 7.5e7 mouse brain neurons [Marblestone, 2013]. One of the considered approaches was recording cation changes into replicating DNA by a polymerase that misincorporates bases at higher calcium ion concentrations. They evaluated each approach based on spatiotemporal resolution and power dissipation, noting such a "molecular ticker tape" has a distinct advantage of inherently enabling single-cell resolution and low energy consumption. However, more engineering is necessary on the speed and synchronization of the polymerase for this method to be realized [Zamft, 2012].

Synthetic Nucleic Acid Based Memory

To asses the novelty of this thesis, it is necessary to review other prior work that has implemented synthetic forms of nucleic acid based memory storage.

In 2000, Gardner et al demonstrated a genetic circuit with stable toggle switching between two expression states. The switch was implemented using two inducible genes negatively regulating one another by expressing a repressor of the other gene's promoter [Gardner, 2000]. Recent work by Pam Silver used a two gene circuit for implementing a latch switch. In this case, one inducible "trigger" gene expressed the inducer for the "latch" gene, which also expressed the inducer for itself to achieve autoregulation. In both toggle and latch examples the state of the system is propagated by RNA transcription and can be stable for several days [Kotula, 2014].

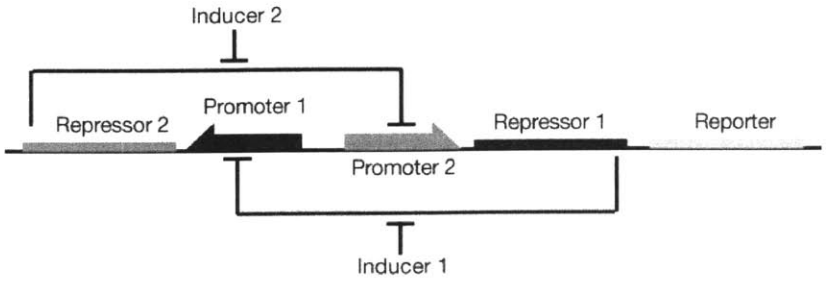


Figure 3. A genetic toggle switch by Gardner et al. Courtesy of Nature.

Modifications to DNA propagated by replication provide a more stable form of memory. Several groups have done so to implement reversible bits, logic gates, counters, barcode generators, and more using a class of proteins called site specific recombinases. Such recombinases, like Cre, Flp, and Rci, join DNA at recognition sequences associated with the recombinase and can then integrate, translocate, remove, or flip the DNA segments based on the relative orientations of the recognition sequences. Since the recombinase can join almost any pair of its recognition sequence, all aforementioned recombinase-based memory implementations restrict the number of any recognition sequence to two copies [Yang, 2014].

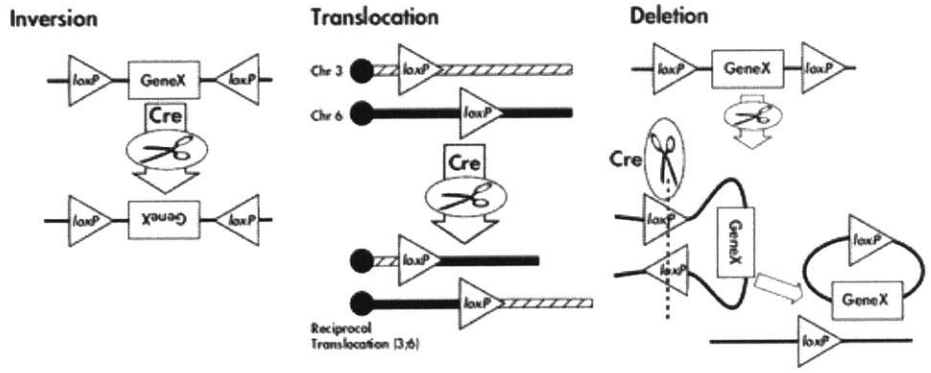


Figure 4. Recombination outcomes of Cre and its LoxP recognition sequence. Courtesy of JAX Cre Repository.

The amount of memory that can be achieved with any of the current approaches is

strictly limited by the number of inducers, repressors, and recombinases that can work together without cross-talk. Effort has been placed on expanding this number by means of directed evolution on proteins or mining genomic databases for new parts, but neither has led to significant scalability in the size of the genome that can be made available for data recording. Note, in preparation of this thesis, new work demonstrates the potential combine reverse transcription and lamda red recombination to store programmable memory in an *E. coli* chromosome [Farzadfard, 2014].

Directed Endonucleases

This thesis takes a new approach to recording information into the genome by use of a group of proteins, known as directed endonucleases, which have programmable specificity to DNA sequences [Esvelt, 2013]. Programmable specificity allows the library of available recognition sequences to grow exponentially with the length of the recognition sequence. Scalable memory is achieved by the short mutation a programmable directed endonuclease can leave on its recognition sequence.

This process starts when the directed endonuclease localizes on its DNA recognition sequence, where it then procedes to cleave both strands of DNA. The ends of such double-stranded breaks (DSBs) are quickly repaired by native end joining processes, such as Non-Homologous End Joining (NHEJ) or Alternative End Joining (AEJ). which often insert or delete bases from the DNA ends. These mutations are called "indels" and modify a recognition sequence site in a way that prevents future localization of the directed endonuclease. So if each recognition sequence is viewed as a binary digit, then each directed endonuclease-caused indel can be viewed as an irreversible bit flip.

Examples of directed endonucleases include Zinc Finger Nucleases (ZFNs), Transcription Activator Like Effector Nucleases (TALENs), and some proteins associated with Clustered Regularly Interspaced Palindromic Repeats (CRISPR), like Cas9 and Cascade. Zinc Fingers (ZFs) and Transcription Activator Like Effectors

(TALEs) are proteins with peptide sequences designed as a succession of fused protein domains that each contribute recognition to a subsequence of the overall DNA recognition sequence. ZFNs and TALENs are made from ZFs and TALEs by additionally fusing a DNA-cleaving domain, like FokI. Alternatively, site specificity through the CRISPR operon is guided by an RNA molecule that forms a RNA-protein complex with Cas9 [Jinek, 2012]. Part of the guide RNA sequence codes for affinity to the protein and another part is a 20 bp sequence (known as "spacer") that enables Cas9 to localize on DNA complimentary to this stretch of the RNA (a recognition sequence also called "protospacer") as long as the recognition sequence is adjacent to a short pattern of DNA (the "protospacer adjacent motif" or PAM). Cas9 naturally contains the necessary domains to cleave DNA. The RNA-based programmability of the CRISPR system has had a tremendously rapid and constructive impact in scalable genome engineering, genetic screens, DNA-encoded logic, gene expression control, antimicrobials and more.

Genomically Recorded Accumulated Memory (GeRAM)

To overcome the limitations of previous approaches to nucleic acid based memory, this thesis demonstrates a cell based Genome Recorded Accumulated Memory (GeRAM). Data from cell based sensors is recorded via Cas9 and guide RNA into repetitive non-coding regions of a mammalian genome and into a fluorescent reporter gene on bacterial cell plasmids. GeRAM data is later read out by sequencing the recognition sites and inferring the degree of exposure to the inducing signal by a comparison to an uncut reference genome.

As shown in the following figure, a pathway dependent promoter is used for inducible expression of guide RNA and a constitutive promoter for Cas9. This approach allows multiple guide RNA constructs with different site specificity to direct Cas9 to numerous recognition sites, lending further scalability in the number of pathways simultaneously writing to GeRAM.

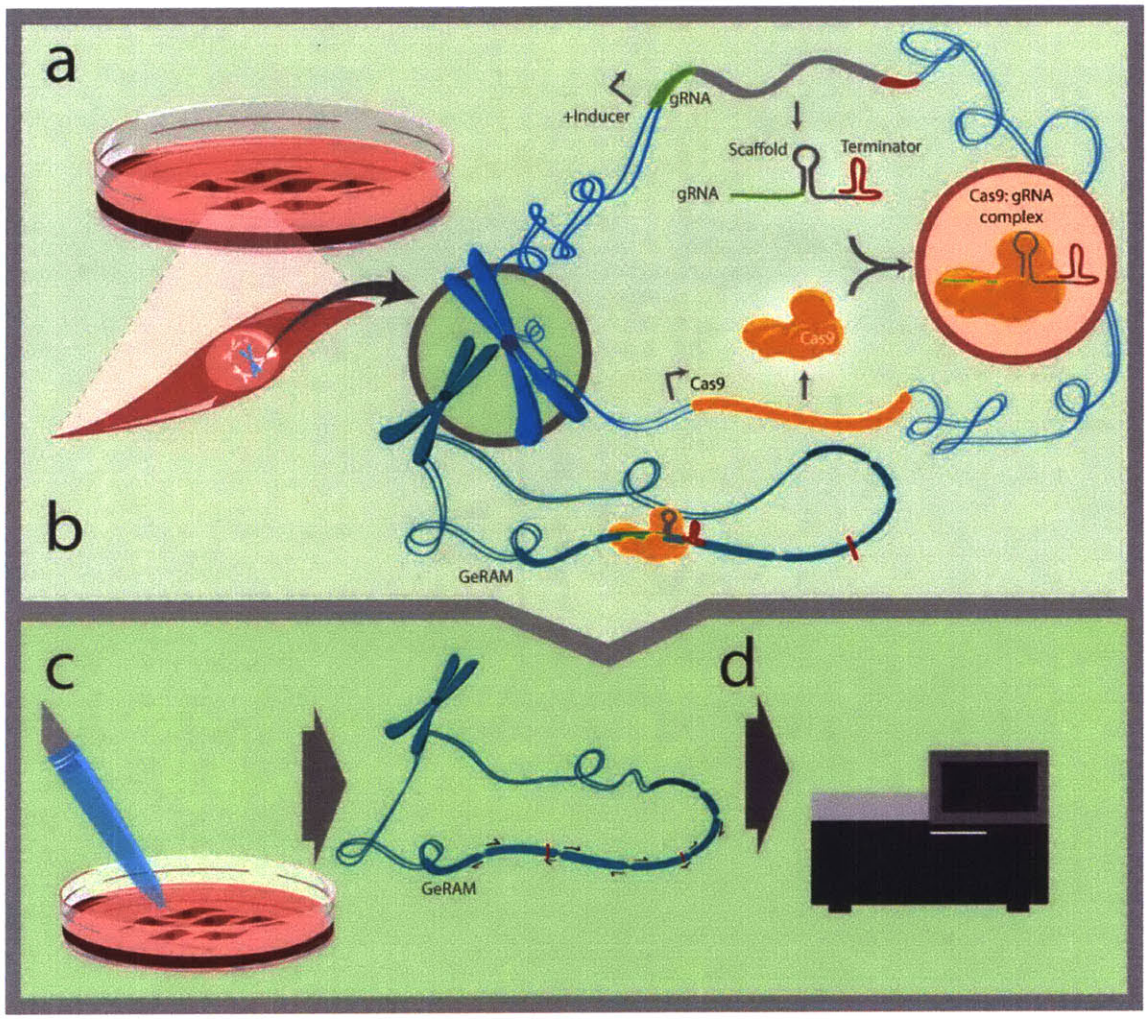


Figure 5. Genome Recorded Accumulated Memory workflow illustrated by Lisa Nip. a) Cell culture express Cas9 constitutively and gRNA by induction. Cas9 protein recognizes and complexes with induced gRNA to target cleavage of GeRAM; b) Coexpression of gRNA and Cas9 results in accumulation of indel mutation memory in GeRAM; c) Cells are harvested and GeRAM in amplified by polymerase chain reaction (PCR); d) Amplified DNA products are sequenced.

Chapter 2:
GeRAM in Bacterial Culture

"Anything found to be true of E. coli must also be true of elephants."

- Jacques Monod, 1954

Overview

We cloned Cas9 and guide RNA into BL21 *E. Coli*, a strain that naturally lacks these CRISPR components [Jiang, 2013], with the goal of causing indel mutations on a single DNA recognition sequence (one bit) within the gene for a fluorescent protein reporter. Early results are shown that suggest a correlation in the duration of chemically induced expression of guide RNA and the fraction of fluorescence lost within a bacterial population. The chapter concludes with a working protocol for the assembly of DNA with a high density of repeated guide RNA recognition sequence (multi-bit). The experiments in this chapter were conducted in large part with Lisa Nip.

Experimental Setup

The two plasmid constructs transformed into *E. Coli* are illustrated in Figure 6. One plasmid contains Cas9 derived from *S. pyogenes* and a chloremphenicol drug resistance gene, both under constitutive (constantly active) expression. The second plasmid contains guide RNA driven by an arabinose sugar inducible promoter (pBAD), as well as the genes for green flourescent protein (GFP) and ampicillin/ carbenicllin resistance under constitutive expression. The guide RNA's spacer sequence (GAGAAATACTAGATGCGTAA) targets GFP within the first 100 bases.

To achieve inducible expression, *E. Coli* were electroporated with both plasmids, outgrown for one hour without antibiotic, and then grown in a minimal nutrient growth media containing 0.01% arabinose plus 1000x carbenicllin and chrolemphenicol.

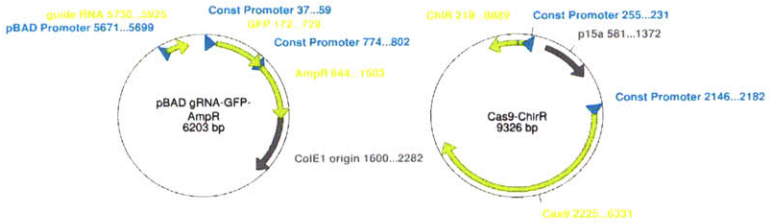


Figure 6. Plasmids used for an Implementation of GeRAM in Bacteria

Chemical Induction of Insertion and Deletion (Indel) Mutations

After overnight growth, plasmids were harvested from *E. coli* and Sanger sequenced. Figure 7 illustrates an overlay of the aligned ab1 trace files for samples sequenced with a forward primer shortly upstream of GFP. Within this excerpt of the trace files there is noticeable dephasing four bases upstream of PAM. The observed position of dephasing is consistent with the pattern of indels left by Cas9-guide RNA complexes in a high-copy number collection of plasmids from *E. Coli* [Jiang, 2013].

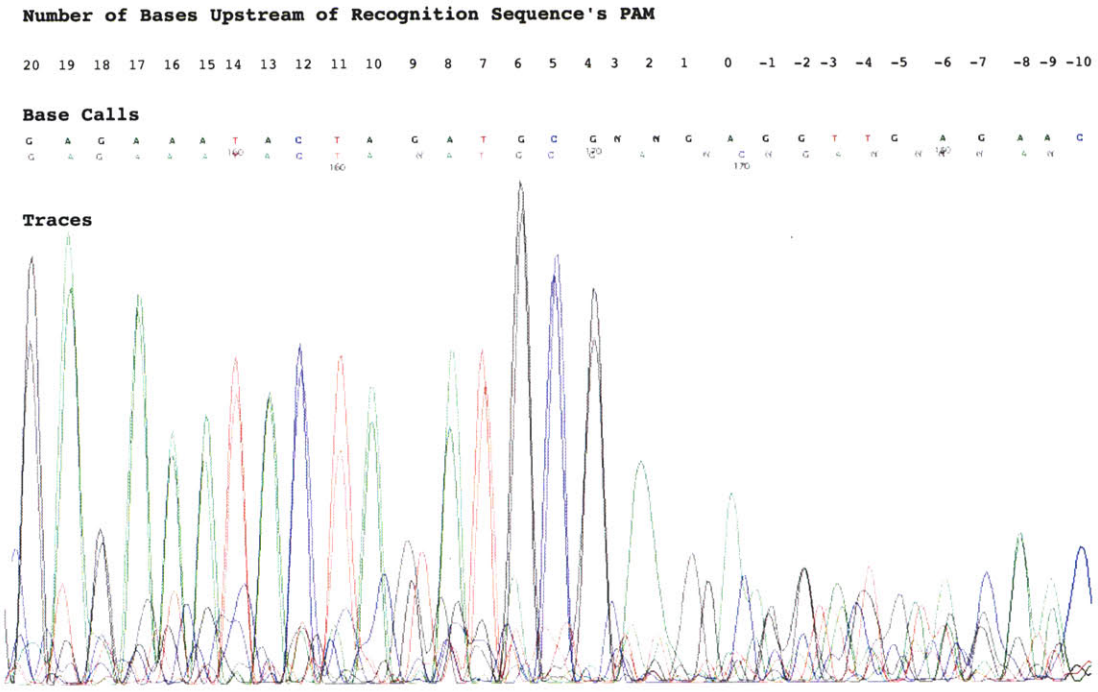


Figure 7. Overlay of trace files indicating an indel mutation in the recognition sequence.

In a following experiment using the same constructs, *E. coli* were cultured in the same minimal growth media supplemented with 0.01% arabinose, but for variable durations (0, 5, 10, and 15 minutes). After each duration of arabinose exposure, cells were plated on a 10 cm Petrie dish containing solid minimal media and grown for an additional 15 hours. The resulting colonies were checked for GFP expression and imaged using a UV transilluminator, which emits within GFP's absorption spectra. Colonies that did not fluoresce most likely expressed a non-functional GFP transcript containing a frame-shifting indel mutation within the guide RNA's recognition sequence. As seen in Figure 8, longer exposre to arabinose resulted in a greater fraction of the colonies that lost fluorescence.

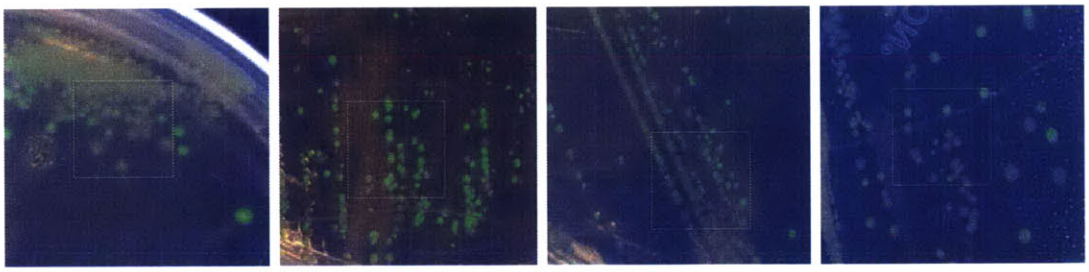


Figure 8. Observed loss of fluorescence correlating with arabinose exposure.

Duration of Chemical Induction (minutes)			
0	5	10	15
<u>Within Yellow Region of Interest</u>			
# Green Flourescent Colonies			
~40	25	22	3
# Non-Flourescent Colonies			
7	9	~40	15
% Green Flourescent Colonies			
85	73	55	16

<u>Assembly of Synthetic GeRAM Sequences</u>

The *E. coli* genome is limited in repeat sequences and therefore also limited in the size of naturally available GeRAM. To overcome this limitation, a synthetic insert with multiple recognition sequences was constructed using pairs of DNA oligos. Four of the oligo pairs contain the recognition sequence and a two nucleotide identification. When

annealed, these four oligo pairs have two 3' overhanging ends that enable the pairs to ligate together in any random order. The random ordering effectively barcodes subsequences of the assembly and can later on facilitate precise mapping of sequencing data back onto the entire molecule. Two other terminal oligo pairs only contain one 3' overhanging end, which prevents ligation from extending the molecule at the other end.

As shown in Figure 9, synthetic GeRAM containing up to 11 repeats was assembled by mixing the interior oligo pairs at 5x and the terminal oligo pairs at 100x in a ligation reaction. Assemblies were amplified by polymerase chain reaction, size selected by electrophoresis on an agarose gel, isolated by topoisomerase cloning into DH5a *E. Coli*, and lastly confirmed by Sanger sequencing. Rolling-circle replication can be tested as an alternative synthesis method.

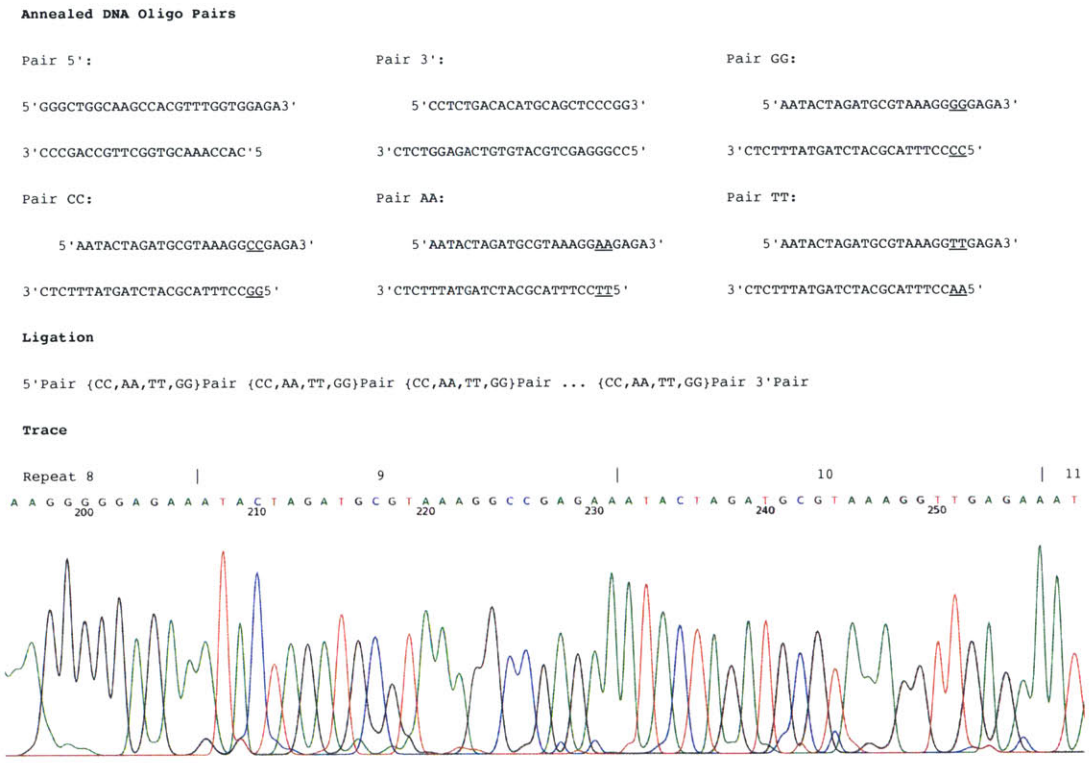


Figure 9. Top: DNA oligo pairs for synthetic GeRAM assembly. Bottom: Trace file showing the 8th, 9th, 10^{th,} , and 11th repeat of a sequenced assembly.

Chapter 3:
GeRAM in Mammalian Cell Lines

"As my students and postdocs have often remarked to me, it is easy to generate outlandish results with the PCR but difficult to show that they are correct"

- Svante Paabo, Neanderthal Man: In Search of Lost Genomes, 2014

Overview

This chapter covers recording data into multiple identical recognition sequences (multi-bit) within a mammalian genome. Towards this end, we packaged CRISPR guide RNA and Cas9 into lentiviruses and infected those into an immortalized cell line of mouse embryonic fibroblasts (MEFs). We selected recognition sequence from the mouse genome's tandem repeats, stretches of the genome with a consecutively occurring pattern in DNA sequence. The first part of this chapter covers the algorithm for selecting ideal tandem repeats from a genomic database. Early experimental results suggest a correlation in the duration of constant guide RNA expression and the observed percentage of indels within repeats. The experiments in this chapter were conducted in large part with Dr. Naama Kanarek.

Selection of Natural GeRAM Sequences

Mammalian genomes are replete with repeated sequences. Repeated sequences interspersed throughout all chromosomes make up 46% of the 3.3e9 base pair human genome and 38% of the 2.8e9 base pair mouse genome. Since bit storage is limited by the number of 23 bp sequences containing the necessary NGG PAM sequence, the upper bound on the number of bits available to GeRAM using natural non-coding repeat sequence is 66 Megabits for the human genome and 46 Megabits for the mouse genome.

However, such unchecked pervasive shredding and repairing of DNA would undoubtedly have a drastic cost on cell viability. Therefore, repeats chosen for natural GeRAM are instead selected from a class of repeats that occur in tandem rather than interspersed.

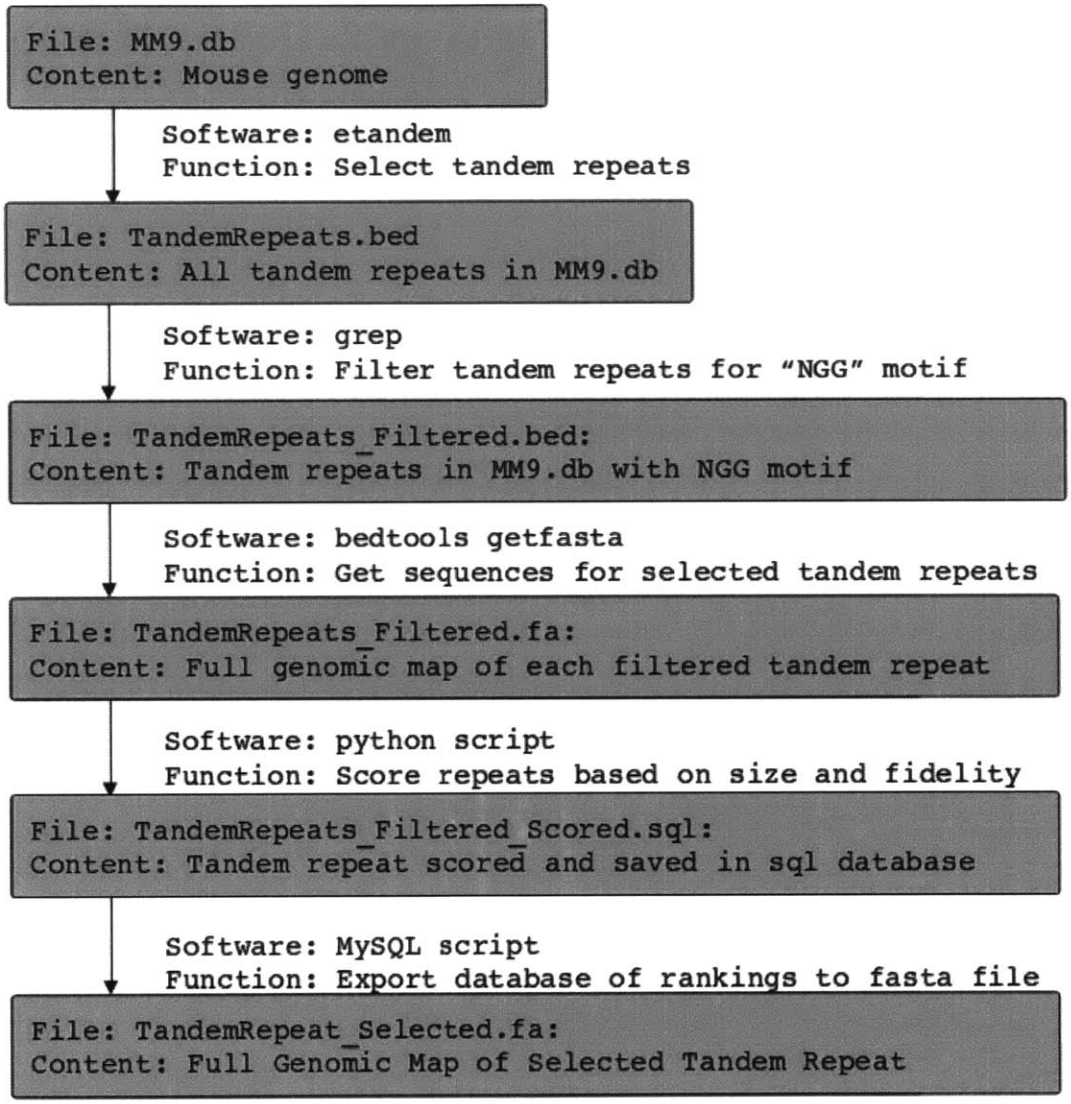


Figure 10. Outlined workflow for tandem repeat selection showing processes (next to arrows) operating on files (gray boxes) .

Tandem repeats make up 2.5% of the mouse genome. The procedure designed to select tandem repeats for GeRAM is outlined in Figure 10. Standard bioinformatics and shell command line tools were used to compile all tandem repeats from the mouse genome containing GG or CC. Repeating sequences less than 20 basepairs long were discarded because cuts on these smaller repeats are more likely to result in the removal of entire repeats by microhomology end joining repair. Next, a custom python script generated a MySQL database of tandem repeats passing the filter with scores on their number of repeats, number of repeats with sequence variation, number of repeats with sequence variation on the recognition sequence, and number of repeats with sequence variation on the PAM. Lastly, five candidates were selected through sorting the remaining tandem repeats by their maximum repeat count and ratio of off-recognition sequence variation to on-recogntion sequence variation.

Experimental Setup

The following experiments focused on a tandem repeat from chromosome 7 with 81 counts of a 37 bp repeat according to the MM9 mouse genome assembly. The complete tandem repeat is summarized in Figure 11.

GCTCTGTATAAGGAGCCTTGGAAGGTGCATGCTGGAA(1x)GCTCTGTATAAGGAGCCTTGG
AAGGTGTATGCTGGAA(1x)GCTCTGTATAAGGAGCCTTGGAAGGTGCATGCTGGAA(6x)G
CTCTGTATAAGGAGCCTCGGAAGGTGCATGCTGGAA(1x)GCTCTGTATAAGGAGCCTTGGA
AGGTGCATGCTGGAA(1x)GCTCTGTATAAGGAGCCTCGGAAGGTGCATGCTGGAA(21x)G
CTCTGTATAAGGAGCCTCTTAAGGTGCACGTTGTAA(1x)GCTCTGTATAAGGAGCCTCGGA
AGGTGCATGCTGGAA(45x)GCTCTGTATAAGGAGCCTTGGAAGGTGCATGCTGGAA(4x)

Figure 11. Variants of the selected repeat sequence as ordered within the complete tandem repeat region. Colors indicate repeat variants with the same sequence. Variations from the canonical sequence are in bold. PAMs for the recognition sequence are underlined.

The two lentiviral vector constructs were sequentially infected into MEFs. The first construct contains a human codon optomized Cas9 and a neomycin drug resistance gene, both under constitutive polymerase II expression. After a week-long selection

post-infection of the first construct, surviving MEF cell were infected with the second construct. The second construct contained guide RNA driven either constitutively by a U6 polymerase III promoter or induced by a NF-kappaB promoter (Invivogen) plus the gene for puromycin. After an additional two days of selection, genomic DNA was harvested every other day from cells continuously expressing guide RNA.

Next-Gen Sequencing Sample Prep

Two rounds of PCR amplification were used to prepare a library for each time point and control for 250 bp paired-end read sequencing on an Illumina MiSeq. Both rounds of amplification use Life Technologies Platinum Taq DNA Polymerase along with their recommended concentrations for each part of the reaction mixture.

First Amplification (15 cycles)

```
5'1/2FwdIlluminaAdapter,DiversityTag,_
  ____________________________________|
 |_GCTCTGTATAAGGAGCCTCGG3'
...GCTCTGTATAAGGAGCCTCGGAAGGTGCATGCTGGAA...GCTCTGTATAAGGAGCCTCGGAAGGTGCATGCTGGAA...
...CGAGACATATTCCTCGGAGCCTTCCACGTACGACCTT...CGAGACATATTCCTCGGAGCCTTCCACGTACGACCTT...
                                                          3'TTCCACGTACGACCTT_
                                                  ____________________________|
                                               |_,DiversityTag,1/2RevIlluminaAdapter5'
```

Size Selection (Inset) and Fragment Analysis

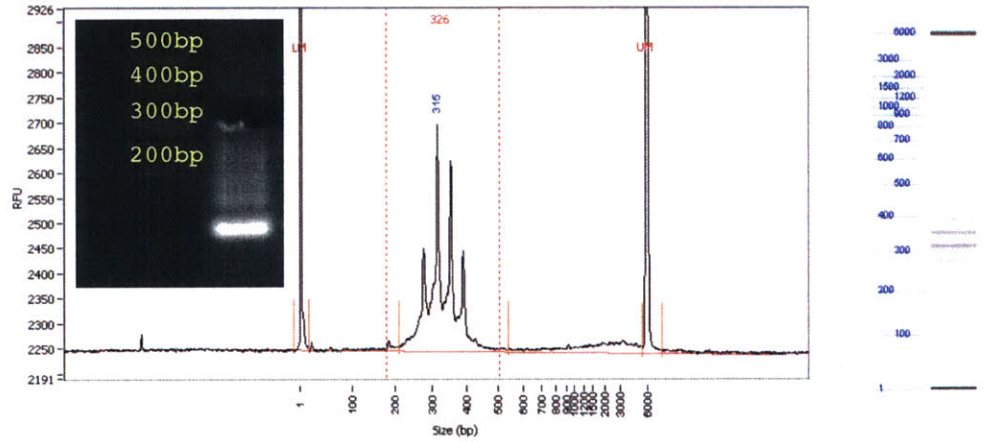


Figure 12. Top: First round amplification primer design. Bottom Inset: Agarose gel imaged after size selection. Bottom: Fragment analysis on the size selected extract.

The first round amplifies internally on the tandem repeat such that differences in amplicon sizes are multiples of the unit repeat length. Primers for the first round have two additional sequences on the 5' end. Preceding the part of the primer derived from the tandem repeat is one of four 8 bp sequences that adds diversity to the sequencing run. This diversity tag in turn improves the overall quality of basecalls. On the very 5' end of the first round primers is one half of the Illumina adapter, which is used in the second round of amplification for initial priming. Accordingly, primers for the second round of amplification are simply the full Illumina adapters, including a barcode sequence for each library.

After both rounds of amplification, the library's reaction was run individually by electrophoresis on an agarose gel and the region between 300-400 bp was extracted. DNA purified from the extract from the first and second reaction were used as input to the second reaction and pooled to a 10uM mixture for sequencing, respectively.

Second Amplification (35 cycles)

```
5'2/2FwdIlluminaAdapter,_
  ____________________________|
 |_1/2FwdIlluminaAdapter3'
 5'1/2FwdIlluminaAdapterDiversityTagGCTC...GGAADiversityTag1/2RevIlluminaAdapter3'
 3'1/2FwdIlluminaAdapterDiversityTagCGAG...CCTTDiversityTag1/2RevIlluminaAdapter5'
                                                        3'1/2RevIlluminaAdapter_
                                                  ____________________________|
                                                 |_,LibraryTag,2/2RevIlluminaAdapter5'
```

Fragment Analysis

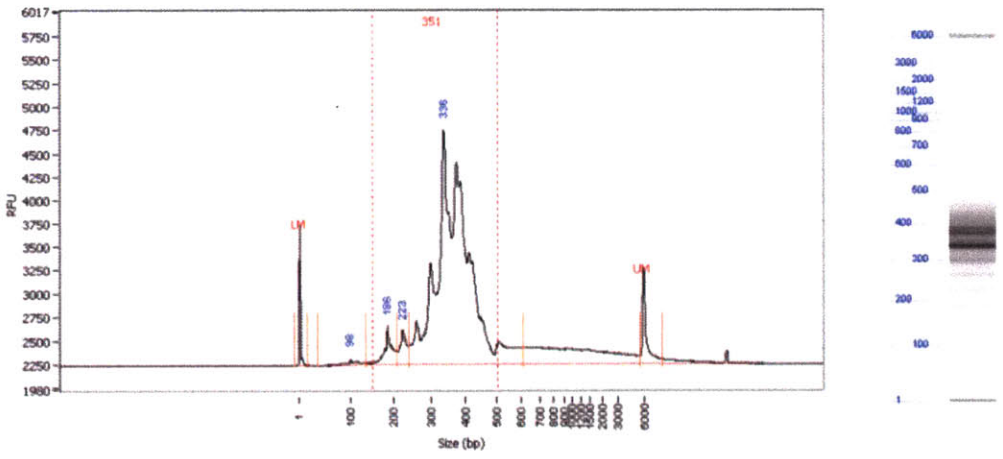


Figure 13. Top: Second round amplification primer design and fragment analysis.

Analysis for Quantifying Indels

The number of reads represented in each library ranged from 2,000 to 350,000. In preparation for analysis two pre-processing steps were applied to the data. Using the fastx-clipper command from the FASTX bioinformatics toolkit, the first step was to "clip" away Illumina adapter sequences from the beginning and ends of reads and filter out any reads shorter than the length of two repeats. The second step discarded reads that did not begin with the forward primer sequence (with the Linux "grep" command). These pre-processing steps ensure that reads passing through both steps represent amplification of genomic material rather than dimerization of the primers.

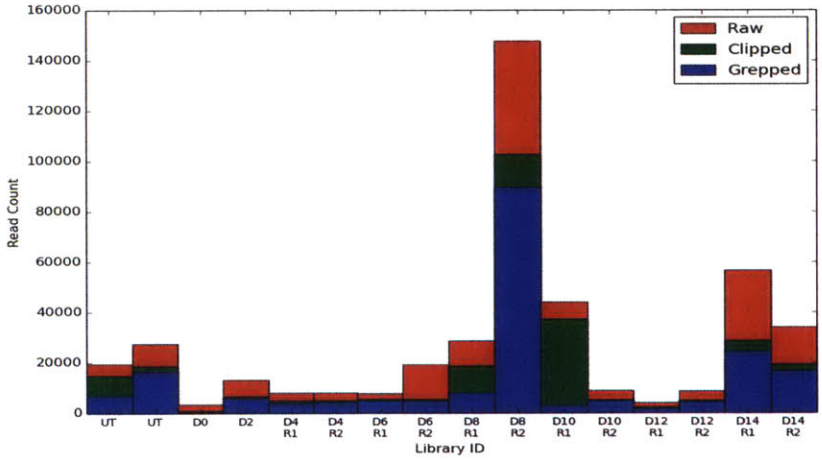


Figure 14. Chart of overlaid read counts. The library ID "UT" indicates control libraries from cells only transduced with the guide RNA plasmid. The library ID "D# R#" indicates the number of days post-selection cells were harvested for the library and the biological replicate ID.

Indels were quantified for all libraries in the sequencing data by matching a sliding window along the reads with one or two lookup tables. The first lookup table consists of the expected sequence of the window if the repeats contained in the window exactly matched wild type repeat sequences. If the match between the window and one of these sequences exceeded 80% – a threshold found to indicate unlikely misalignment - no

indel is recorded at this position and the window is incremented. Otherwise the window is matched against a second lookup table consisting of expected sequences for a range of indel lengths at each position in the window. Indel size and position were then inferred by the best match between the window and lookup table. However, if the best match was below the specified threshold of 80%, then analysis on the read would halt instead of sliding further. As a result of pre-processing the initial position along the tandem repeat is known. The sliding window's position is updated with each increment according to the size and position of the indel identified in its previous position. Scripts for this algorithm can be downloaded from http://web.mit.edu/~njakimo/GeRAM.zip.

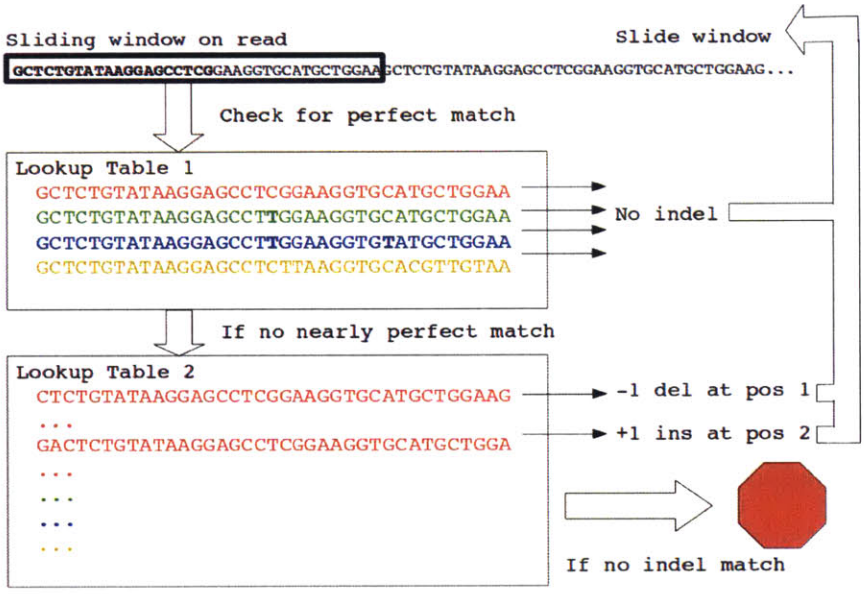


Figure 15. Algorithm for finding indels.

The described algorithm was implemented in python scripts. Shell commands described in Figure 16 were added to manage parallel work flow control on the Whitehead Institutes's computing cluster.

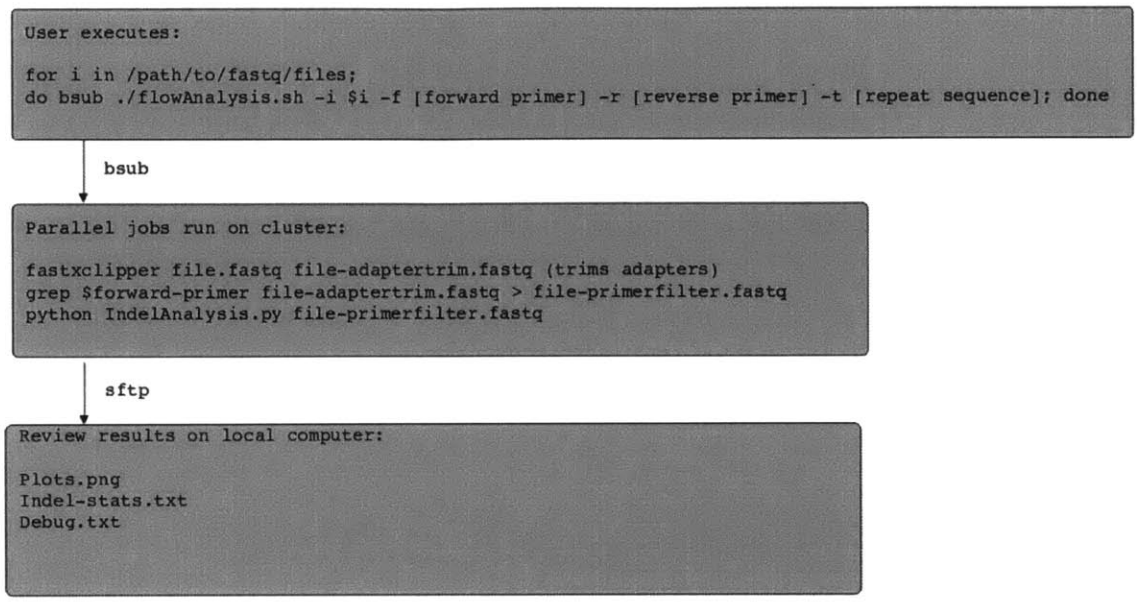


Figure 16. Work flow of indel analysis. Arrows indicate software used for simple parallelization of tasks.

Discussion of Results

The size and position distribution of indels recorded for each library agree with those from related literature [Mali, 2013]. Thus, the results shown in Figure 17 confirm the expected pattern of on-target cutting along the tandem repeat.

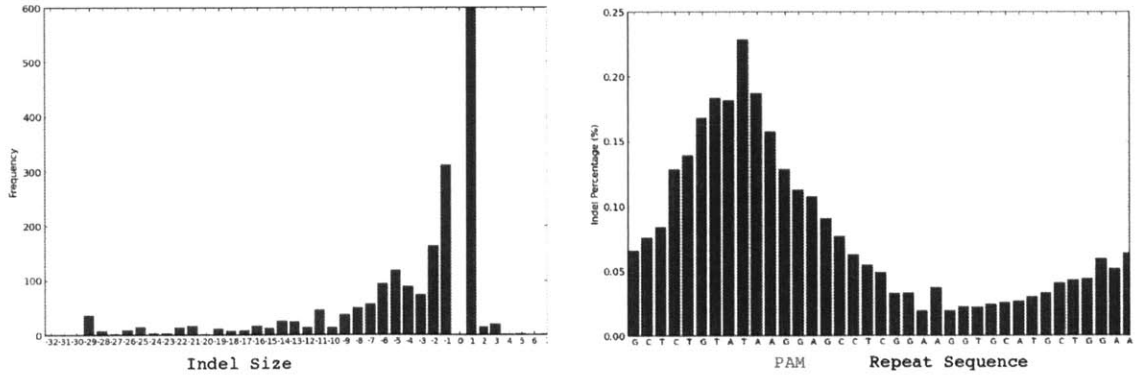


Figure 17. Algorithm for finding indels.

Data shown in Figure 18 illustrates that in a time course with constant induction of guide RNA we observe a consistent increase in the number of indels identified. However, this increase is not at a consistent rate expected from a Poisson process. Likewise, in the NFY induced data (see Appendix) we observed an increase in indels over time at inconsistent rates and no clear correlation with the duration of induction.

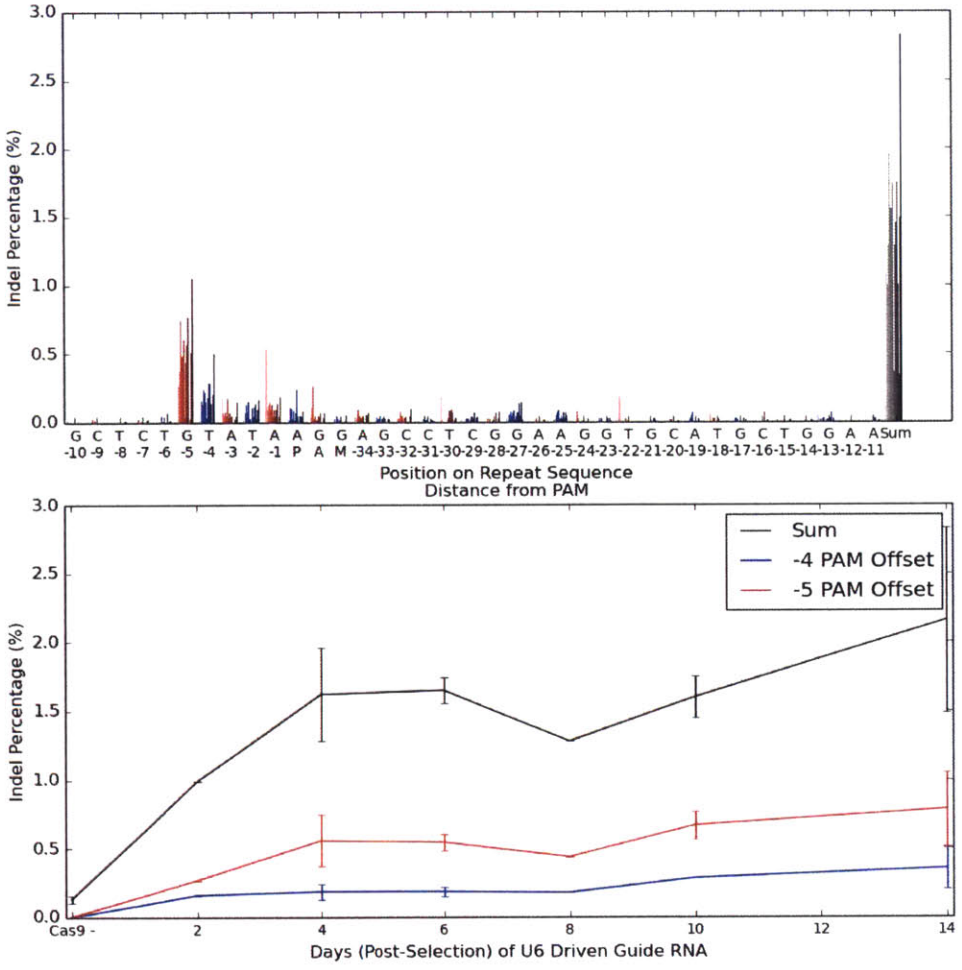


Figure 18. Top: Distribution of indels on repeat sequence over the timecourse. Bottom: Distribution of indels with replicates combined over the timecourse.

Chapter 4:
Future Directions

"I think the message one gets from this was that Linus Pauling was enormously fertile in the way he developed his ideas. He was asked a few years ago, "How do you get ideas?" And he gave I think what is the correct reply. He said, "If you want to have good ideas you must have many ideas. Most of them will be wrong, and what you have to learn is which ones to throw away.""

- Francis Crick, 1995

Overview

Drawbacks of recording information on long tandem repeats include difficulty to detect how sequencing results are influenced by recombinant PCR products during amplification or deletions longer than the repeat sequence during genomic recording. The latter can result from either simultaneous DNA cuts or homologous recombination repair. The first future direction could overcome this challenge by limiting targeted cuts to a single genomic locus that is offset from a recognition sequence. Dispensing with cutting altogether, the second future direction leverages work engineering Cas9 for transcriptional regulation to implement a counting genetic circuit based on programmable activation and repression of genes [Konermann, 2013; Lebar, 2014].

Cuts Offset from Recognition Sequence

By replacing wt Cas9 in our system with one deactivated Cas9 fused to two FokI domains, we can potentially make dense recordings at a single position in the genome. The method could work by repeatedly cleaving a target nucleic acid as a a repeatable directed endonuclease (RDE), wherein the RDE binds to its recognition sequence on the target nucleic acid and cleaves the target nucleic acid at a position that is offset from the

RDE's recognition sequence. The recognition sequence is thefore preserved for the RDE which to make a second cleavage of the target nucleic acid sequence at a position that is offset from the RDE's recognition sequence.

Using this method two or more relocalization events of the targeting domain of the RDE cause additional removal of bases adjacent to the recognition sequence. As illustrated in Figure 19, placement of a gene that expresses a protein or nucleic acid component of the RDE complex within the target nucleic acid can allow this process to terminate once the deleted area extends into said gene. If essential DNA is included in the detectable region of the target nucleic acid, then this process can be used to program a delayed cell death with limited cost to fitness before the deleted region extends into essential DNA. Similarly, non-essential functional DNA can be added to the detectable region for sequential control of a biochemical pathways.

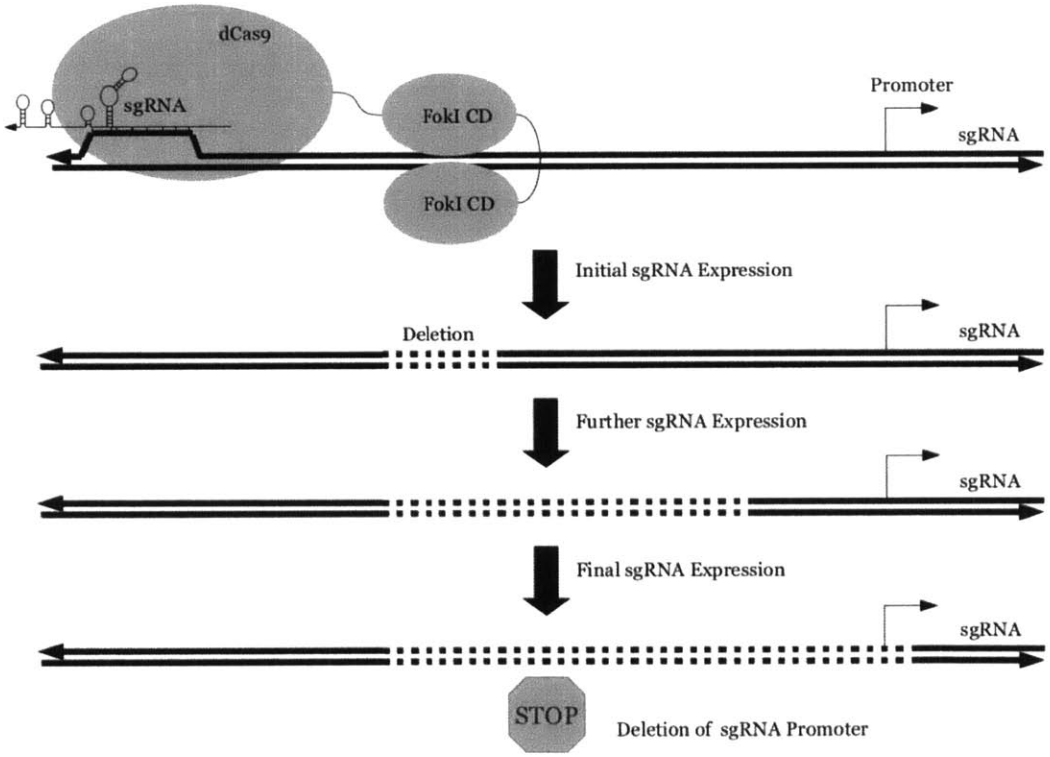


Figure 19. Repeatable directed endonuclease for dense genomic recording.

Feedback Loop

Orthogonally active Cas9 enables a dynamic counter that uses one Cas9 fusion to maintain a memory of the state of the counter and orthogonal Cas9 fusions to increment the state of the counter. Figure 20 shows a schematic representation of a modulo-4 cascaded counter which is clocked (incremented) by light or other induction signal. Application of Cre initiates or resets the counter by flipping the promoter between expression of the first spacer (S) and repeat (R) pair to the expression of an engineered Cas9 that interacts with the transcript of said pair. The state of the counter is preserved with positive feedback on the current expression pattern and negative feedback on the previous expression pattern using spacer and repeat sequences corresponding to this Cas9. An orthogonal Cas9 is either constitutively expressed and complexed with an effector domain upon exposure to light (as shown) or conditionally expressed using an inducible promoter and fused with an effector domain (not shown). Activation or expression of this second Cas9 increments the state using spacer and repeat sequences corresponding to the second Cas9.

The detailed operation of the modulo-4 counter is as follows. Two orthogonal dCas9 ('d' for deactivated cutting domains) are expressed as fusion proteins. dCas9(1) has constitutive expression switched on and off by a recombinase (shown with Cre) and the other dCas9(2) is expressed or activated only upon a signal induction event (e.g. small molecule, metabolite, protein, light, pH etc). When dCas9(1) is not expressed, a guide RNA it complexes with is instead expressed. When dCas9(1) is initially expressed, recent transcripts of such guide RNA initiates the modulo-4 counter by binding to Cas9(1) and the guide RNA's recognition site. This interaction is represented with the black arrow. The position of this recognition site before a weak promoter and linking of dCas9(1) to an effector protein results in the transcription of the "0" state element.

Each element transcribes guide RNA that similarly targets a recognition site before the weak promoter of that element. This positive auto-regulation is represented with a blue

arrows that loops back on the transcriptional memory state element. Each state element transcribes guide RNA that targets a region right before the weak promoter on the next state element in the sequence. However, this guide RNA only complexes with dCas9(2), which is expressed only upon signal induction. Thus, the the counter increments upon signal induction events. This positive feed-forward interaction is represented by red arrows. Each state element also transcribes guide RNA that targets a region right after the weak promoter on the previous state element in the sequence. This guide RNA complexes with constitutively expressed dCas9(1) and will thereby repress the auto-regulated transcription of the previous state. This negative feedback interaction is represented by the blue bar-ended lines.

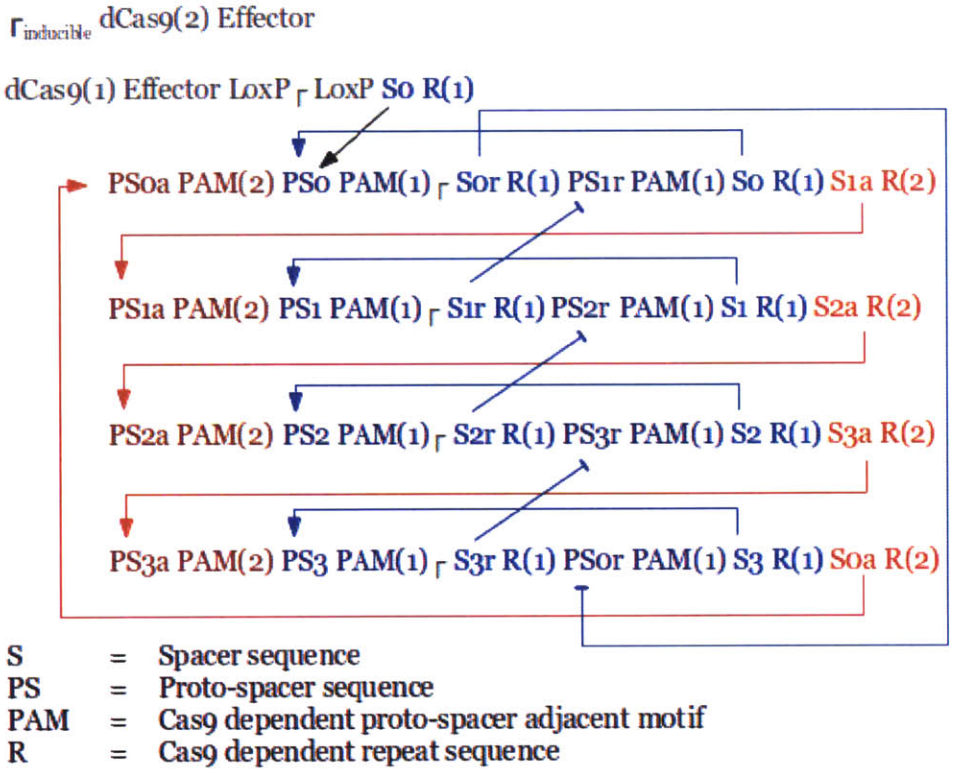
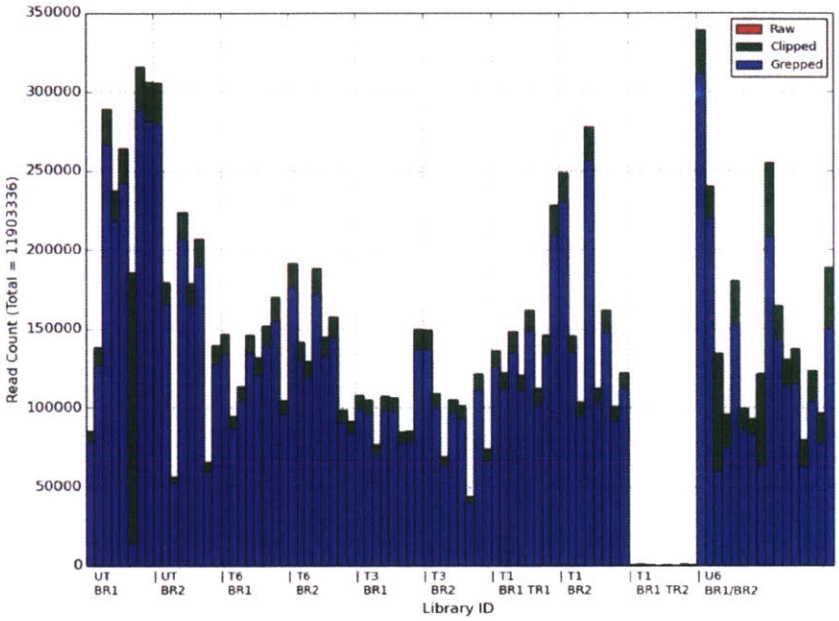


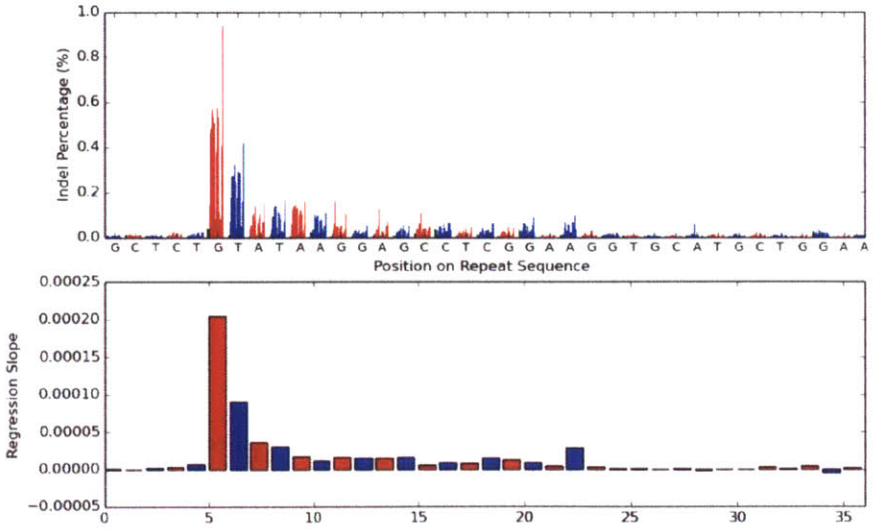
Figure 20. Schematic representation of a dynamic cascaded counter induced by light or other induction signal Cascaded counter induced by light or other induction signal. The dynamic modulo-4 counter increments transcriptional state upon each induction. A recombinase (shown with Cre) can be used to initiate and reset the cycle.

Appendix

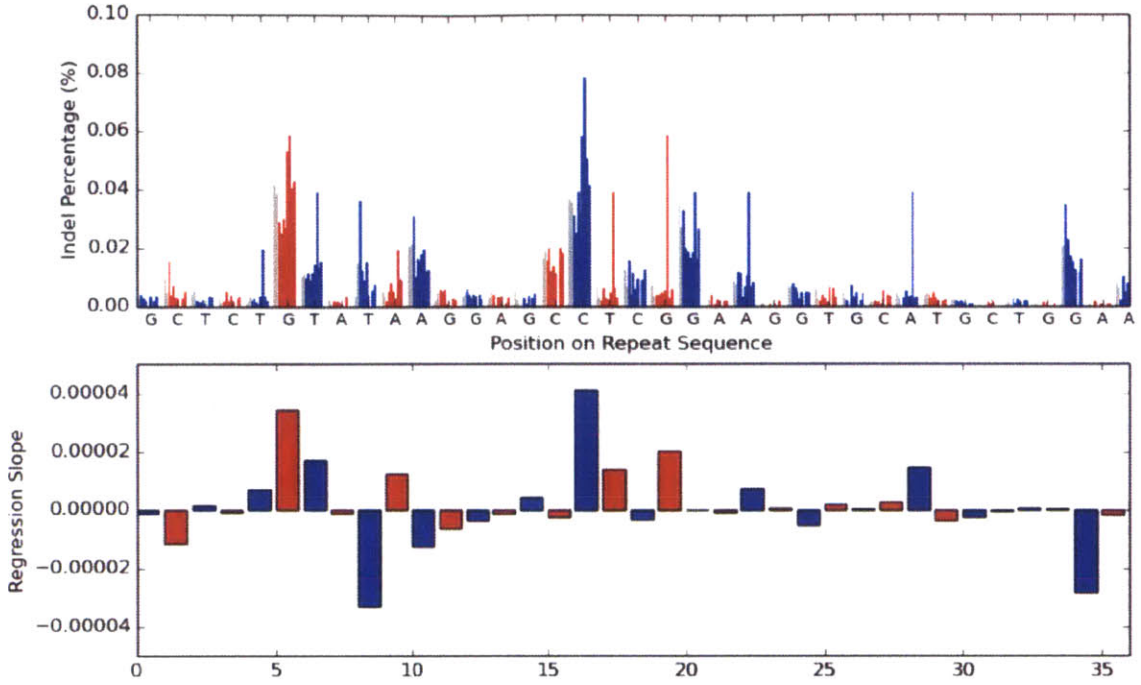
Reads Per Library



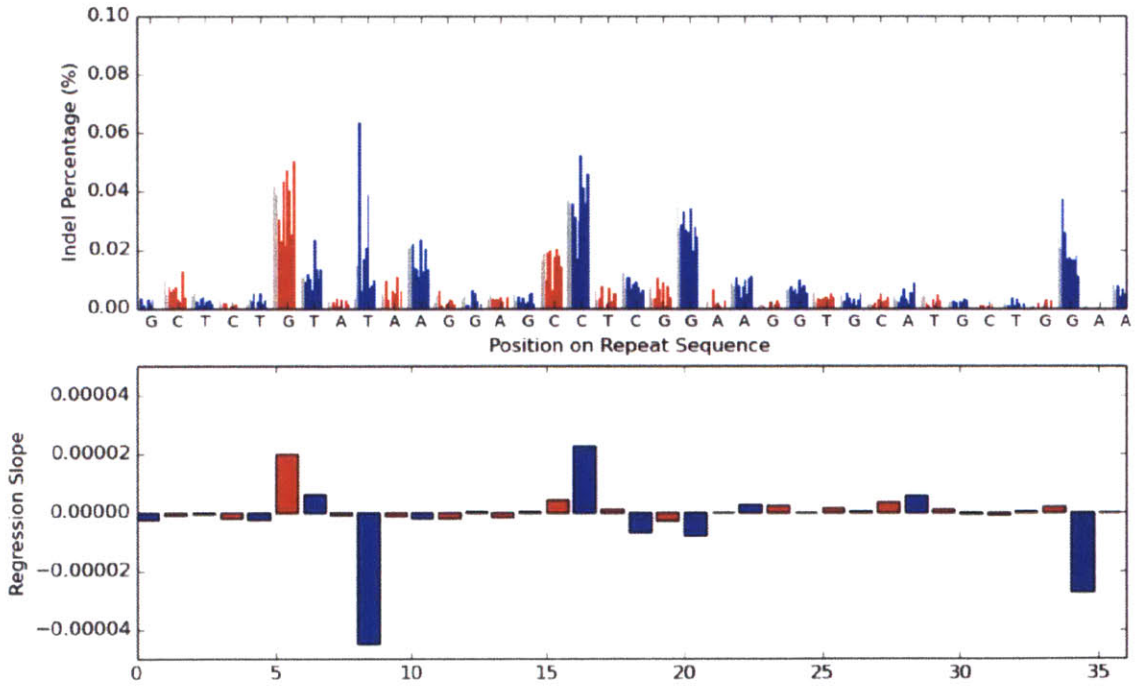
U6Promoter Timecourse Indels



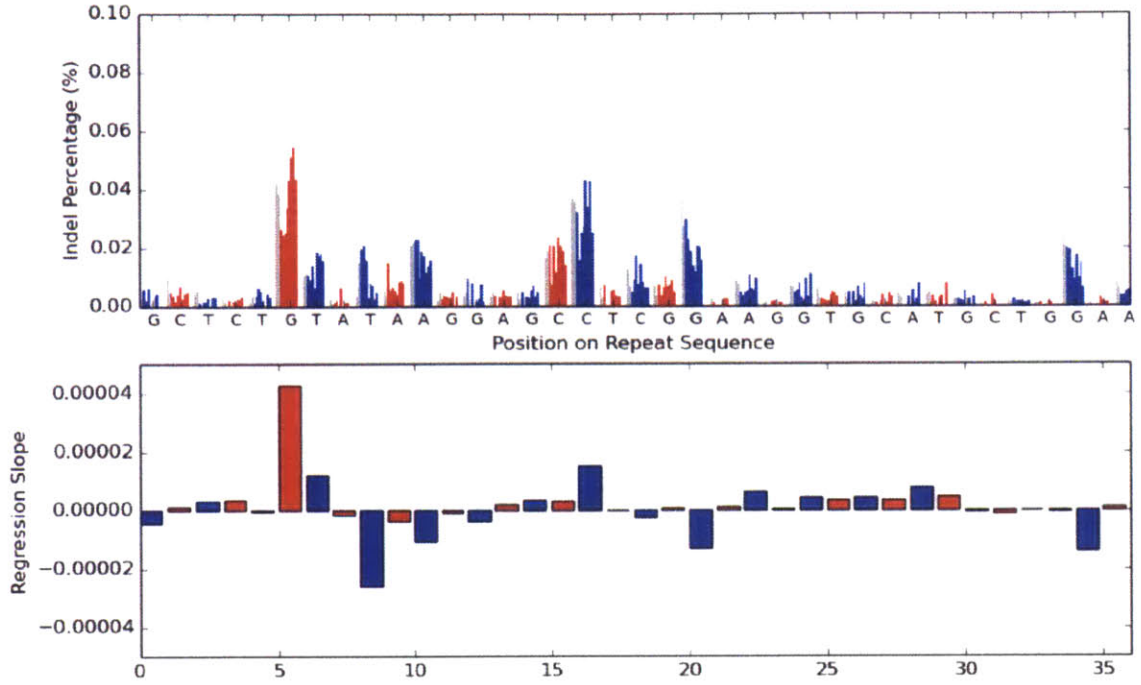
NFY Timecourse Indels (Untreated R1) Results from wt in gray, experimental in color.



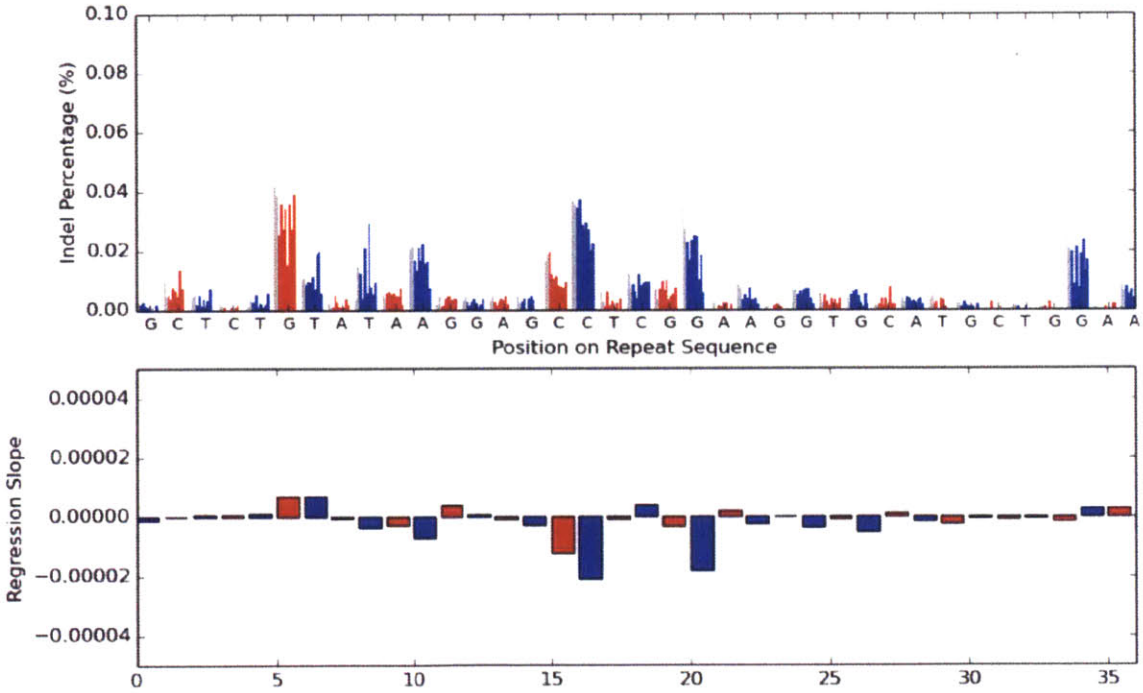
NFY Timecourse Indels (Untreated R2)



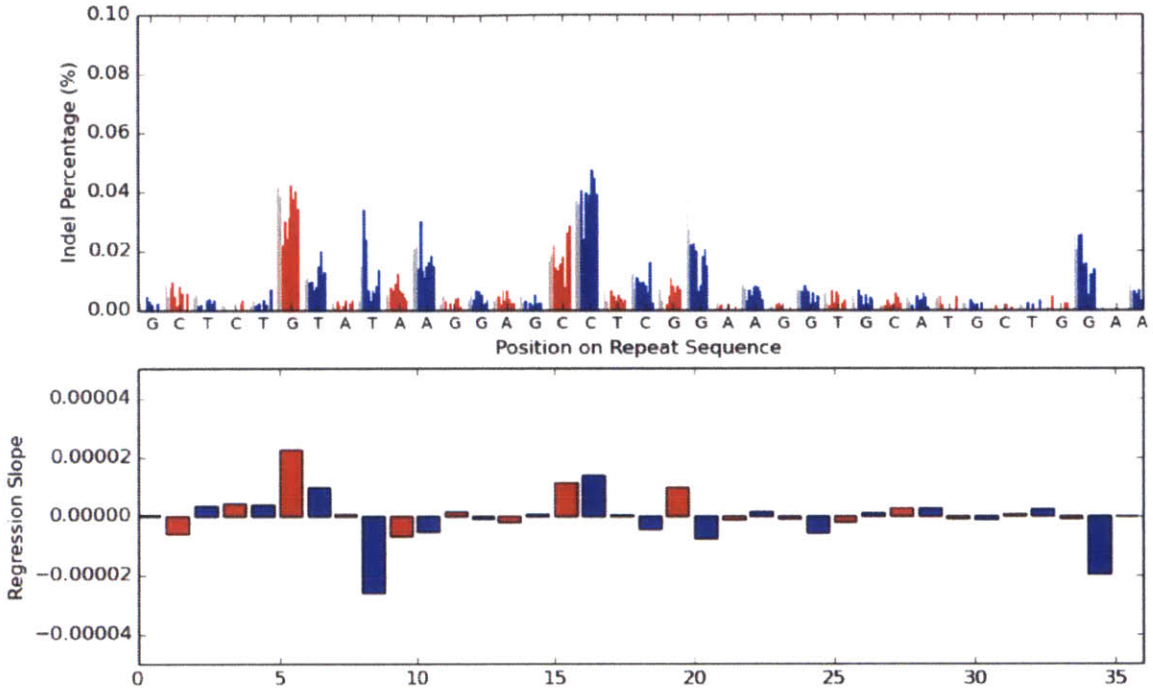
NFY Timecourse Indels (1 Treatment / 6 Day R1)



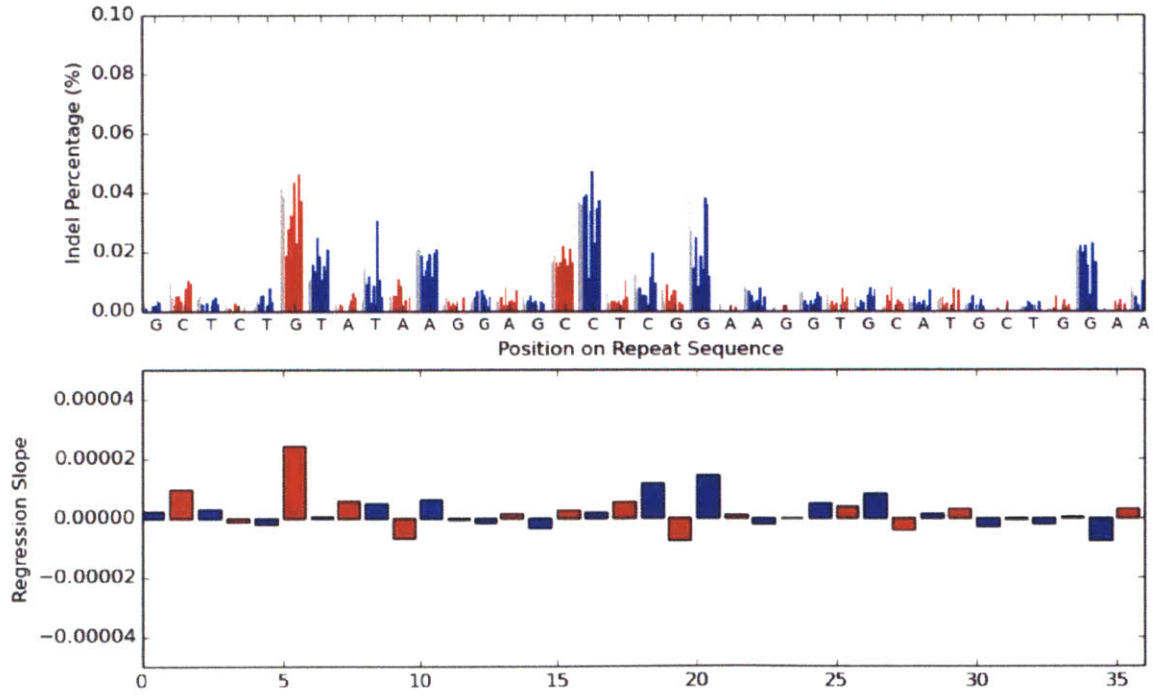
NFY Timecourse Indels (1 Treatment / 6 Day R2)



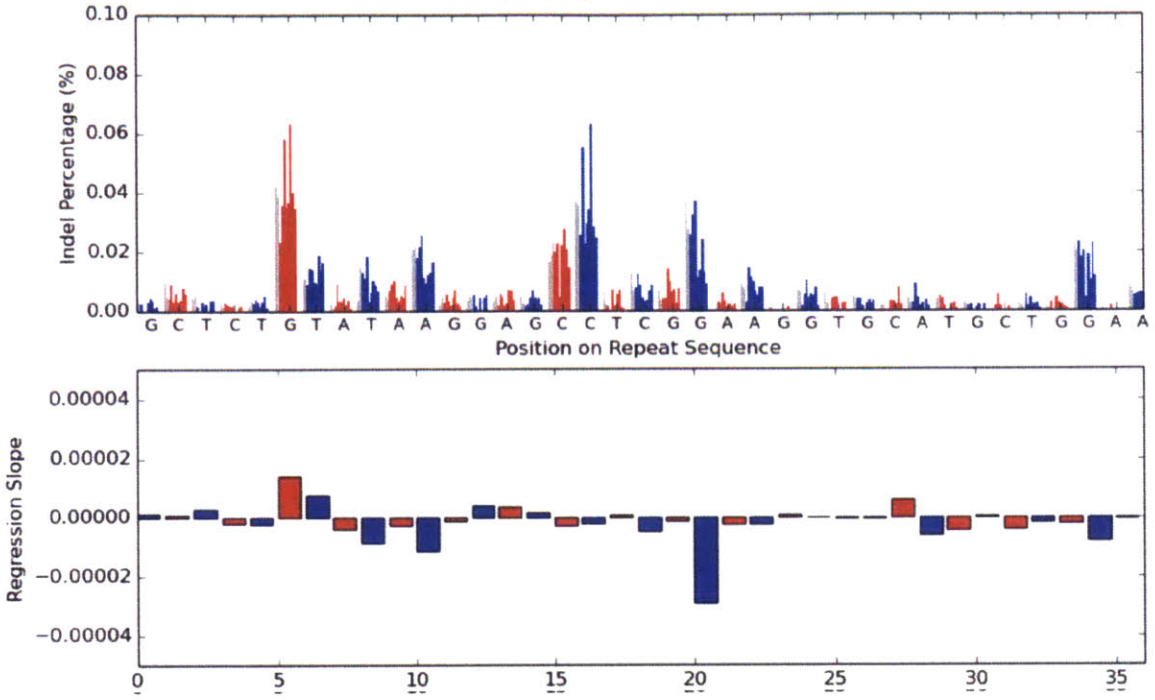
NFY Timecourse Indels (1 Treatment / 3 Day R1)



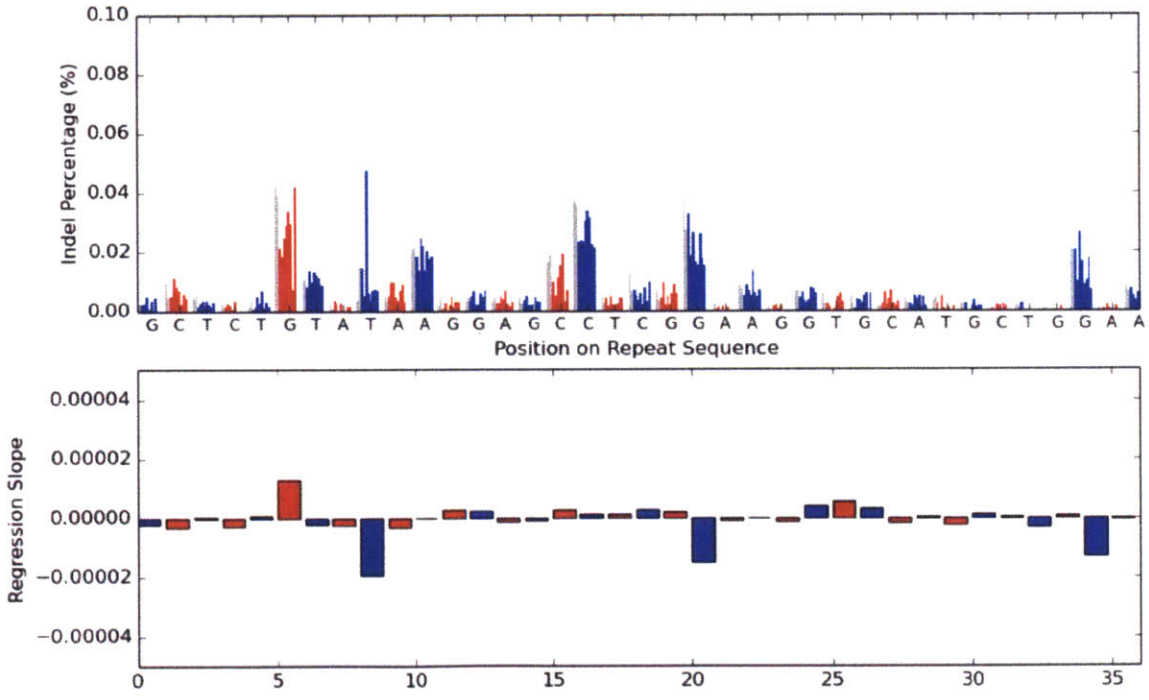
NFY Timecourse Indels (1 Treatment / 3 Day R2)



NFY Timecourse Indels (1 Treatment / 1 Day R1)



NFY Timecourse Indels (1 Treatment / 1 Day R2)



References

Carlson, R. (2014), 'Time for New DNA Synthesis and Sequencing Cost Curves', http://www.synthesis.cc/2014/02/time-for-new-cost-curves-2014.html.

Digikey (2011), 'Online Catalog', http://www.digikey.com/US2011/digikey.pdf.

Esvelt, K. M. & Wang, H. H. (2013), 'Genome-scale engineering for systems and synthetic biology.', *Mol Syst Biol* **9**, 641.

Farzadfard, F. & Lu, T. K. (2014), 'Synthetic biology. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations.', *Science* **346**(6211), 1256272.

Gardner, T. S.; Cantor, C. R. & Collins, J. J. (2000), 'Construction of a genetic toggle switch in Escherichia coli.', *Nature* **403**(6767), 339--342.

Gillings, M. R. & Westoby, M. (2014), 'DNA technology and evolution of the Central Dogma', *Trends in Ecology & Evolution* **29**(1), 1вЂ“2.

Jiang, W.; Bikard, D.; Cox, D.; Zhang, F. & Marraffini, L. A. (2013), 'RNA-guided editing of bacterial genomes using CRISPR-Cas systems.', *Nat Biotechnol* **31**(3), 233--239.

Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J. A. & Charpentier, E. (2012), 'A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity.', *Science* **337**(6096), 816--821.

Konermann, S.; Brigham, M. D.; Trevino, A. E.; Hsu, P. D.; Heidenreich, M.; Cong, L.; Platt, R. J.; Scott, D. A.; Church, G. M. & Zhang, F. (2013), 'Optical control of mammalian endogenous transcription and epigenetic states.', *Nature* **500**(7463), 472--476.

Kotula, J. W.; Kerns, S. J.; Shaket, L. A.; Siraj, L.; Collins, J. J.; Way, J. C. & Silver, P. A. (2014), 'Programmable bacteria detect and record an environmental signal in the mammalian gut.', *Proc Natl Acad Sci U S A* **111**(13), 4838--4843.

Lebar, T.; Bezeljak, U.; Golob, A.; Jerala, M.; Kadunc, L.; Pirs, B.; Strazar, M.; Vucko, D.; Zupancic, U.; Bencina, M.; Forstneric, V.; Gaber, R.; Lonzarić, J.; Majerle, A.; Oblak, A.; Smole, A. & Jerala, R. (2014), 'A bistable genetic switch based on designable DNA-binding domains.', *Nat Commun* **5**, 5007.

Mali, P.; Yang, L.; Esvelt, K. M.; Aach, J.; Guell, M.; DiCarlo, J. E.; Norville, J. E. & Church, G. M. (2013), 'RNA-guided human genome engineering via Cas9.', *Science*

339(6121), 823--826.

Marblestone, A. H.; Zamft, B. M.; Maguire, Y. G.; Shapiro, M. G.; Cybulski, T. R.; Glaser, J. I.; Amodei, D.; Stranges, P. B.; Kalhor, R.; Dalrymple, D. A. & et al. (2013), 'Physical principles for scalable neural recording', *Frontiers in Computational Neuroscience* **7**.

NewEnglandBioLabs (2014), 'SOC Outgrowth Medium - Datasheet', https://www.neb.com/~/media/Catalog/All-Products/913C5621FEF44435AF31DF78AF45CC38/Datacards or Manuals/B9020Datasheet-Lot2191203.pdf.

Phillips, R.; Kondev, J.; Theriot, J. & Garcia, H. (2012), *Physical Biology of the Cell*, Garland Science.

ThemoScientific (2010), 'Heratherm Compact Microbiological Incubator - Product Manual', http://www.thermoscientific.com/content/dam/tfs/LPG/LED/LED Documents/Product Manuals & Specifications/Microbiological Incubators and Environmental Chambers/Microbiological Incubators/D21453~.pdf.

USDepartmentofLabor (2014), 'U.S. city average - Electricity per KWH'.

Yang, L.; Nielsen, A. A. K.; Fernandez-Rodriguez, J.; McClune, C. J.; Laub, M. T.; Lu, T. K. & Voigt, C. A. (2014), 'Permanent genetic memory with >1-byte capacity.', *Nat Methods* **11**(12), 1261--1266.

Zamft, B. M.; Marblestone, A. H.; Kording, K.; Schmidt, D.; Martin-Alarcon, D.; Tyo, K.; Boyden, E. S. & Church, G. (2012), 'Measuring cation dependent DNA polymerase fidelity landscapes by deep sequencing.', *PLoS One* **7**(8), e43876.