Molecular Multiplexing Methods for Genome-Scale Measurements

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Nagendra Prasad Palani

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisor: Peter Tiffin

June 2018

Acknowledgements

I thank Dr. Igor Libourel for giving me an opportunity to pursue my doctoral education in his lab. I thank him for giving me the freedom to pursue my own projects and for instilling quantitative thinking in me.

I thank Dr. Peter Tiffin, who agreed to serve as interim advisor until my graduation. His support has been crucial for me to complete this dissertation.

I thank my committee members – Drs. Jeffrey Gralnick, Kevin Silverstein, Fumiaki Katagiri for their support during my time in the graduate program. I also thank past committee members Drs. Kenneth Beckman and Romas Kazlauskas for their invaluable suggestions and support.

I thank Drs. Kenneth Beckman and Daryl Gohl for constantly reminding me to complete this dissertation while I have been working at the University of Minnesota Genomics Center. I thank Drs. Steve Bowden and Daryl Gohl for their critical reading of this dissertation and their many comments to improve it.

I thank Dr. Brett Barney for his advice and support throughout my graduate education in Minnesota.

I thank my lab mates Dr. Steve Bowden, Dr. Hong Yang, Dr. Eli Krumholz, David Burdge, Joshua Goldford, and others who were instrumental in me gaining a well-rounded education.

I thank the Plant Biological Sciences (PBS) directors of graduate studies Drs. Cindy Tong and Sue Gibson for their advice, support, and helping me stay on track. I thank the PBS graduate students with whom I had an opportunity to interact and have fun with.

I thank the employees at the core facilities (UMGC, UIC, CMSP) for their support of my projects. I also thank PBS and MnDrive for their fellowships that financially supported the bulk of my doctoral education.

I thank my mother Bhavani and my sister Sivaranjani for their support during my pursuit of a Ph.D. I cannot thank enough my wife Ajeetha, who has somehow managed to be my bedrock of support while pursuing her own Ph.D.

Dedication

I dedicate this dissertation to all my teachers,
with special thanks to Ms. Jayanthi.

Abstract

In **Chapter 1**, I present an overview of concepts that underlie the projects discussed in this dissertation, primarily the utility of unique DNA barcodes to tag distinct genotypes and subsequently link them to phenotypes. Such molecular tagging allowed us to perform multiplexed phenotype analysis of thousands of genotypes using next-generation sequencing (NGS) technologies.

In **Chapter 2**, I present *2D Tn-Seq*, a massively multiplexed experimental approach to interrogate genetic interactions of a microbe at the genome scale. We developed 2D-Tn, a synthetic nested transposon that created two orthogonal dimensions of transposon insertions in *Escherichia coli*, with the insertion positions linked by a DNA barcode. A library of Tn5 primary mutants (1st dimension) was pooled and the mariner transposons nested within the primary insertions were induced to transpose *in vivo* and create a secondary transposon insertion (2nd dimension) independent of the primary mutant location. Because each primary mutant can give rise to several secondary mutants, 2D Tn-Seq enabled rapid generation of nearly 100 million double mutant lineages within a few weeks. To demonstrate a proof-of-concept that secondary insertion locations could be associated with their primary insertion locations, a library of approximately 10^4 secondary mutants was created from a pool of approximately 10^4 primary mutants. The primary and secondary insertion locations, and the linking barcodes were identified by short-read NGS. Orthogonal insertions that must be present in the same cell were linked by the shared DNA barcode. Current iteration of the bioinformatic analysis could associate approximately 7% of the secondary insertions to their primary insertion locations, demonstrating that the molecular steps work as designed. We expect the method, which was conceived to be portable to most eubacteria, to democratize acquisition of exhaustive genetic interaction datasets.

In **Chapter 3**, I describe the development of a molecular tool that will enable peptide-based ^{13}C metabolic flux analysis (MFA) of a mixed population of microbial cells. The plasmid pFluxSeq is a synthetic DNA vector designed to express DNA barcodes as heterologous recombinant proteins in *E. coli* strains. The plasmid enabled selection of translatable DNA barcodes that result in stable and functional proteins. Recombinant barcode proteins could be separated at high purity from the native proteome by dual protein purification tags, suitable for discerning isotope labeling patterns by mass spectrometry. A set of *E. coli* Keio deletion collection mutants that are informative of central metabolism were transformed with pFluxSeq to associate their genotypes with unique DNA barcodes. This collection can be pooled and used to determine flux maps of all the constituent genotypes simultaneously using peptide-based MFA.

In **Chapter 4**, I describe the application of deep mutational scanning to arrayed protein libraries. The mCherry fluorescent protein gene was randomly mutated, linked to DNA barcodes, expressed in *E. coli*, and individual colonies of the mutant library were arrayed in 96-well plates to build an ordered collection. From a single pool of the library, the genotypes contained within the collection were identified *en masse* by single molecule real time long read NGS. Genotype-linked DNA barcodes were associated to the spatial locations of the mutants by orthogonal sample pooling and short read NGS, thus associating individual genotypes to their positions in the array. Protein fluorescence phenotype was recorded for each mutant by array position, thereby linking mutant genotypes to their respective phenotypes. Based on this linking, we explored how structural properties of mCherry influence its phenotype, discovering novel amino acid positions important for mCherry spectral properties while confirming known mutational hotspots.

In **Chapter 5**, I present an analysis of dCas9 mediated gene expression interference in *E. coli* using a multiplexed CRISPR library that targets coding sequences genome-wide. The biological reproducibility of CRISPR interference screens was investigated through use of DNA barcodes that tag and delineate independent colonies sharing a CRISPR guide RNA genotype. Lineage replicates multiplexed into a single experiment exhibited greater concordance in fitness compared to biological replicates from experiments performed on different days. Further, the efficacy of temporally controlled dCas9 expression was studied in the context of its utility in screening the same CRISPR library across different conditions. Based on fitness analysis of individual lineages of each guide RNA genotype, a case is made for introducing barcodes in otherwise isogenic colonies as a means to observe inter-colony variation in growth studies.

Table of Contents

List of Tables

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Contribution of authors

I performed the work included in this dissertation as the research requirement for the degree of Doctor of Philosophy at the University of Minnesota.

Chapter 2

I independently conceived the project and was performing preliminary experiments when we found out that Dr. Steve Bowden in Prof. Hirotada Mori's lab (NAIST, Japan) was working on a similar idea. We subsequently chose to collaborate.

I performed experiments to test Tn5 transposome electroporation, positive selection post FRT-recombination, and *in vivo* mariner transposon mutagenesis by promoter-trap mobilization. I equally contributed towards making synthetic molecular constructs and transposon mutant libraries. I performed all bioinformatic analyses of NGS data. I wrote the chapter with input from Dr. Steve Bowden.

Chapter 3

Dr. Igor Libourel conceived the project and I developed it into the version presented here. I made all the molecular constructs, performed all experiments except preparation of NGS libraries (done by Dr. Steve Bowden), and analyzed the NGS data. I wrote the chapter.

Chapter 4

I conceived the project and managed the collaboration. I made the molecular constructs, generated the mCherry mutant library, and performed phenotyping. Dr. Daryl Gohl and Archana Deshpande supervised colony picking, orthogonal pooling, and NGS library preparation. Dr. Kenny Beckman mapped the barcodes to the positions. I performed the bioinformatic analysis to map genes to positions using the barcodes. Dr. Igor Libourel performed the linear modeling, generated the mCherry structure-function relationship, and wrote the Results section. I wrote the rest of the chapter.

Chapter 5

I conceived the project, performed all experiments, analyzed the data, and wrote the chapter.

For all of the projects, Dr. Igor Libourel contributed suggestions to improve the projects into the form presented in this dissertation.

Chapter 1

Introduction

Functional genomics drives a systems understanding of biology

Life as embodied by a self-replicating cell does not arise from the individual functions of the constituent cellular components but rather from interactions of those functions with one another and with the external environment. This applies even to the simplest theoretical automaton ancestor that can be described as alive (Ganti 2003). Thus, to understand the behavior of a living cell in its environment, it is necessary to adopt a global view rather than look at the parts in isolation. The systems approach to biology involves developing predictive models of biological phenomena to understand how individual cellular components lead to emergent behaviors, which are complex responses not encoded in the biological parts catalog but which rather arise from the amplitude and frequency of spatio-temporal interactions between those parts (Ideker, Galitski, and Hood 2001). Predictive accuracy of the model is dependent on the extent of our comprehension of the biological system, and it is therefore a requisite that comprehensive information be collected on how the manipulation of each component's function and the interactions of those functions affect the system as a whole. Discrepancies between model prediction and experimental observation is minimized and eventually resolved by an iterative process of refining the model based on hypothesis-driven generation of additional and often new types of functional genomic information (Auffray et al. 2003).

Functional genomics is the generation of genotype-phenotype relationships at the genome scale using high-throughput methods. It provides the knowledge base for model building in systems biology. Functional genomics methods apply the same type of perturbation individually to every component of a system in a given level of organization (eg: inactivating genes in a genome) and measure the resulting effect on a global parameter (eg: fitness of the organism). These methods have been used to acquire data in a highly parallel fashion (i.e. simultaneous measurement of physically separated analytes) and increasingly have transitioned towards incorporating multiplexing as a core property (i.e. simultaneous measurement of pooled analytes). Multiplexing

is desirable because only a single instance of measurement is required to simultaneously collect data for all the analytes, which is more cost-effective than parallel measurements.

Functional genomics methods in bacteria

The need for functional genomics tools in bacteria

Kingdom Bacteria constitutes one of the most abundant, widespread, and persistent classes of organisms on our planet, found even in hostile subglacial lakes (Bulat 2016) and spacecraft clean rooms (Vaishampayan et al. 2013). The importance of bacteria in human health, the environment, and industry cannot be understated. Human bodies contain as many or more microbial cells as mammalian cells (Sender, Fuchs, and Milo 2016). The vast majority of biochemistry encoded within bacterial metabolism is undiscovered and is a key driver of nutrient cycling at the planetary scale. However, beyond a few model organisms like *Escherichia coli*, our knowledge of the functional capabilities of bacterial species drops off precipitously. Even for the model organisms, the availability of functional genomic tools is limited compared to that of eukaryotic models. While we have been cognizant of the need to understand the immense genetic diversity encompassed within the bacterial kingdom, only recently are we developing the functional genomics tools required to explore in high-throughput the biology of non-model organisms (Wetmore et al. 2015). Further, development of tools for bacteria can inform design of similar tools for archaea, a very understudied but ecologically essential domain of life that thrives in extreme environments and shares several genome features with bacteria (eg: lack of introns, polycistronic) while more closely related to eukarya in their biochemical apparatus.

High-throughput genotype generation

Creating genome-scale genotype variants is a common starting point for most functional genomics assay. Disrupting a gene sequence to cause a loss of gene function is often the simplest way to create novel genotypes, and its effects are measurable across several phenotypes. For example, disrupting a gene encoding an enzyme relevant in cellular metabolism can alter the rate of substrate to biomass conversion, thus affecting the organism's growth rate which is also reflected in changes in the transcriptome, proteome, metabolome, and fluxome. Most often, genes are inactivated by transposon driven insertional mutagenesis. Transposons are naturally occurring parasitic mobile genetic elements that can insert themselves within a host DNA sequence. Several transposons have been engineered into genetic tools that enable facile and near-random insertional mutagenesis throughout the target genome (Rubin et al. 1999;

Goryshin et al. 2000). Homologous recombination (HR) mediated gene deletion can precisely remove a target sequence from the genome, with HR driven by systems native to the organism or heterologously introduced (Eg: Lambda red recombination (Datsenko and Wanner 2000)). Often, the target gene is replaced with a defined gene cassette for positive selection of the newly created strain. While HR driven gene deletions are considered the gold standard in novel genotype creation, they are also the most laborious to perform at a genome-wide scale, primarily due to the requirement of creating and maintaining each genotype individually. Both transposons and gene replacement cassettes can be engineered to carry genetic payloads of interest like 1) genes for antibiotic resistance (positive selection for target disruption), 2) transcriptional fusions that can identify open reading frames being expressed under a given condition (Eg: β-galactosidase assay), or 3) outward reading promoters (eg: T7) to drive expression of adjoining sequences. More recently, the CRISPR/Cas9 RNA-guided DNA endonuclease system (Jinek et al. 2012) has been developed into a formidable genome engineering tool that provides precision sequence disruption along with multiplexing capability. While it has seen limited use in prokaryotes due to sparsity of native non-homologous end joining (NHEJ) mechanisms, recent developments (Garst et al. 2017) point towards a multiplex-capable system that can combine the precision of gene replacements with the ease of transposon mutagenesis. Apart from loss-of-function genotypes, other functional genomics genotypes include transcriptional repression or activation, inducible ORF libraries (Kitagawa et al. 2005) that can either be plasmid borne or chromosomally integrated (H. H. Wang et al. 2012), and direct RNA interference based on recently discovered Cas13a (Abudayyeh et al. 2017). Within a gene, every single nucleotide can be mutated to other bases to create an exhaustive single amino acid variant library of the encoded protein (Haller et al. 2016).

Molecular barcodes tied to genotypes enable highly multiplexed phenotyping by next-generation sequencing

Cost-effective scalability is a highly desired property in any functional genomics method because greater the information we can collect about a biological system, the better our representative model of that system. A central innovation in increasing the throughput of functional genomics experiments has been the concept of sequence barcodes. Although methods for multiplexed omics measurements exist with widely varying levels of scalability (**Figure 1**), the greatest degree of multiplexing in phenotype measurement can be achieved by taking advantage of the sequence properties of genotype-linked barcodes. This is due to the fact that measurements of nucleotide sequences can be performed using massively parallel sequencing (or NGS - next generation

sequencing) (Shendure et al. 2017), which is one of the most cost-effective and scalable measurement platforms.

Nucleotide barcodes are synthetic sequences of arbitrary length (usually < 50 bases) that are used as unique labels for genotypes. Barcode-tagged sequences can also be thought of as encased in uniquely identifiable virtual compartments. The barcodes can either be random, partially defined or fully defined sequence, and need not encode any biologically relevant information. Barcodes have several properties that provide them with powerful utility in high-throughput experiments (**Figure 2A**). By tagging an arbitrary genotype with a unique barcode, that genotype can be detected in a mixture of genotypes. By counting the number of occurrences of the barcode in a mixture, the abundance of the linked genotype can be estimated. The barcode can also be transcribed and translated so that it is detectable in the transcriptome (Dixit et al. 2016; Jaitin et al. 2016) or the proteome (this work, Chapter 3), thus linking the genotype to these phenotypes. Lastly, barcodes can be generated at extremely high diversity from a simple alphabet, satisfying any complex multiplexing need now or in the future (**Figure 2B**). For example, a random 20 bp DNA barcode synthesized as a degenerate oligonucleotide sequence (all 4 bases equally probable at a given position) encompasses 4^{20} unique sequences that can be used to tag and distinguish more than 10^{12} genotypes, while costing less than $5 to synthesize (from IDT DNA as of April 2018).

Barcodes linked to genotypes can be used as readouts for phenotypes, thus linking genotypes to phenotypes. Use of NGS for such readouts facilitates obtaining highly multiplexed measurements of diverse phenotypes. The power of transferring the phenotype measurement from its original omic domain to the barcode's nucleotide domain can be illustrated by the most commonly performed functional genomics assay - fitness measurement in loss-of-function genotype libraries. Conventional measurement of non-competitive fitness of a mutant genotype relative to another genotype (wildtype or mutant) is performed by growing the genotypes separately in individual culture tubes and enumerating the colony forming units (CFU) at two different time points for each genotype. From the CFU frequencies, relative fitness of a strain can be calculated (as described in (Wiser and Lenski 2015)). Systematic deletion libraries in *E. coli (Otsuka et al. 2015)* and *Saccharomyces cerevisiae* (Giaever et al. 2002) incorporate barcodes within the gene cassettes that replaced non-essential ORFs in the genomes of these organisms. Every genotype in the collection is identifiable by a unique barcode and the entire library is pooled to create a single inoculum. The change in frequency of the barcodes (proxy for genotypes) is measured in the course of a pooled growth experiment by NGS (Bar-seq (Smith et al. 2009)), and the fitness of each genotype is calculated relative to the library or to the wildtype strain (**Figure 3**). Similarly,

Tn-Seq enables *en masse* measurement of organism fitness for a library of transposon insertion mutants (van Opijnen, Bodi, and Camilli 2009). In each mutant, the genome junction adjacent to the transposon acts as a barcode that is uniquely mapped to a single location on the reference genome. Relative fitness values of the mutant genotypes are calculated from the change in frequency of the junction barcodes, as described for Bar-seq. Thus, barcoding genotypes allows us to perform NGS based fitness assays on multiplexed libraries. Other notable applications of genotype-linked barcodes for phenotype measurement include analyses of genetic interactions (Jaffe, Sherlock, and Levy 2017), protein-protein interactions (Yachie et al. 2016; Schlecht et al. 2017), protein structure - function relationship (Fowler et al. 2010), and lineage tracking (Levy et al. 2015).

Challenges in developing functional genomics tools for bacteria

Several challenges exist in implementing new tools just because of the sheer diversity encountered in bacterial species. Assuming we can culture a species of interest axenically in the laboratory, the following are some of the basic challenges one can run into when developing high-throughput tools for the organism.

1. Many species and even strains within a species are recalcitrant to genetic transformation and yield very low number of transformants, whether done through natural competence, electroporation, or bacterial conjugation. The defense systems (CRISPRs, restriction digestion systems) within these species actively recognize and destroy non-self DNA that enters the cell. Often, the first step towards performing extensive genetics in non-model bacteria is to either create or isolate a mutant strain that has increased tolerance towards introduction of foreign DNA.

2. It is difficult to create a library of targeted gene deletion or replacement genotypes because many genome-sequenced species lack NHEJ genome repair systems and homology directed recombination often requires long regions of homology of at least few hundred nucleotides. However, transposon mutagenesis is a good alternative where we sacrifice precision for speed and ease of use in creating loss-of-function genotypes.

3. Some advanced functional genomics methods, like the GIANT-coli system (Typas et al. 2008) to study genetic interactions in *E. coli*, makes use of highly engineered molecular processes that cannot be cost-effectively ported to other species or even to other strains of the same species.

Given the diversity of bacterial flora, tools developed for one species most likely will not directly work in another species and will require customization. Inventing new methods that will report on the same phenotypes as in existing methods but are more amenable to customization will drive expansion of functional genomics studies to lesser studied but important species.

4. Customization of available functional genomics tools to study non-model organisms can be difficult due to the poor performance or unavailability of species-specific equivalents of genetic parts used in constructing the tool (eg: antibiotic resistance genes for positive selection of transformants, plasmid replicons, promoters for heterologous gene expression etc.). Therefore, it is imperative to use genetic parts that tolerate a wide host range and have minimal host dependencies. Mobile genetic elements such as transposons and self-transmissible broad host-range plasmids (Popowska and Krawczyk-Balska 2013) are good sources for such genetic parts.

Extending the palette of functional genomics tools in bacteria was the key motivation in developing the tools and methods described in the forthcoming chapters of this work.

Figures

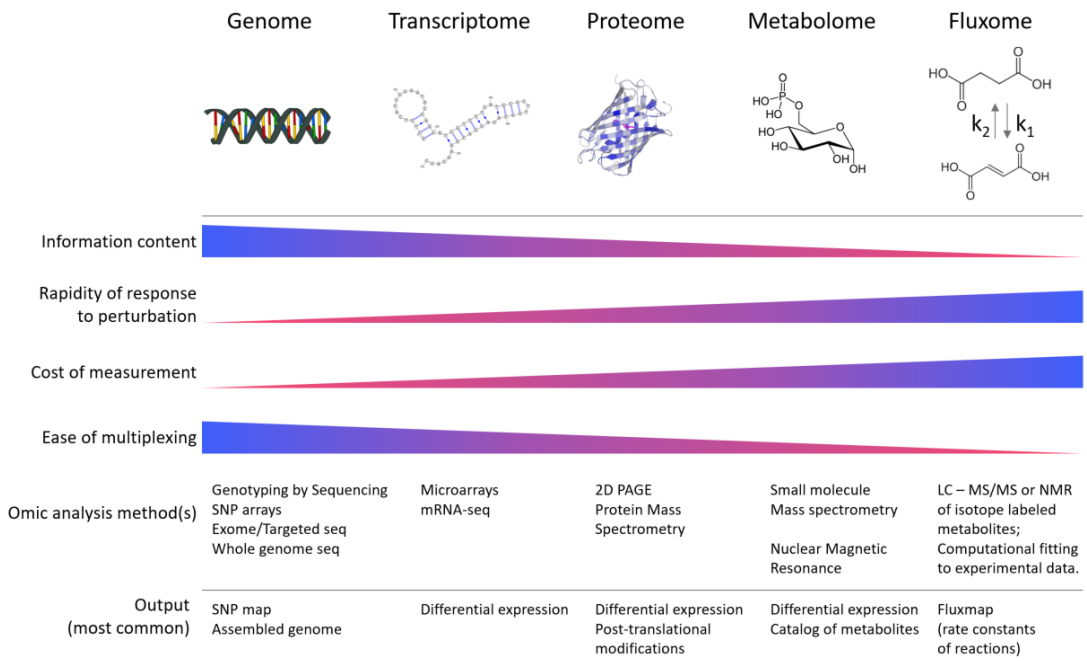


Figure 1: Properties of the primary ‘omes’ relevant to functional genomics studies in bacteria.

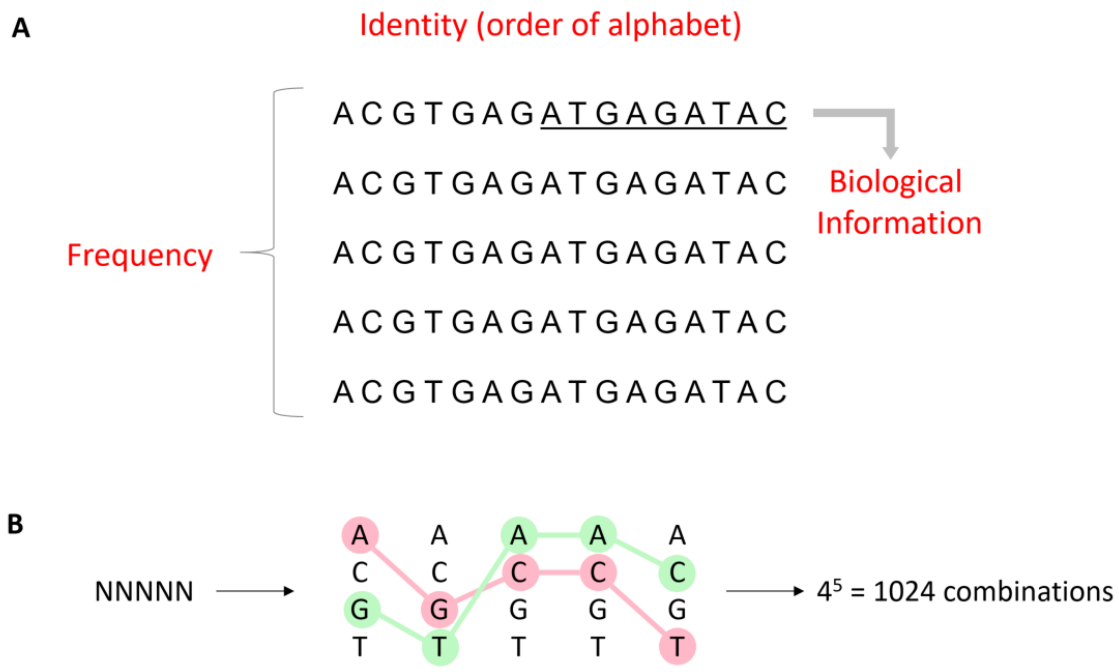


Figure 2: Properties of nucleotide barcode sequences A) Barcodes used as genotype labels are distinguished from one another by their identity, frequency, and encoded content. B) The small alphabet size of canonical DNA nucleotides is advantageous in generating barcode diversity often in excess of current measurement capabilities.

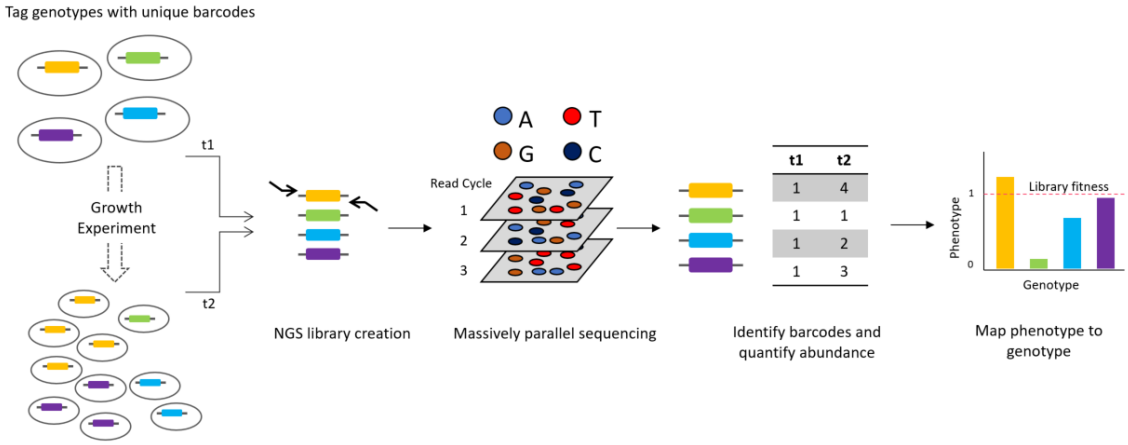


Figure 3: An example for how barcodes enable high-throughput linking of genotype to phenotype. In a simple growth study, several genotypes associated with unique barcodes of known identity are pooled to perform the experiment. Samples are collected over the course of the experiment, subject to NGS library preparation, and 2nd generation sequencing. From the sequencing output, the identity and the abundance of the barcodes are ascertained. From this data, a fitness phenotype is determined for each barcoded strain. Because the barcode - genotype association is known *a priori*, we can link the phenotypes to the genotypes.

Chapter 2

2D Tn-Seq: Multiplexed Genetic Interaction Analysis at the Genome-Scale

Nagendra Palani*, Steven Bowden*, Hirotada Mori, Igor Libourel

*Equal contribution

Introduction

The genes of an organism encode components of a biochemical network which forms the underlying basis of cellular life and reproduction. Because connectivity is a fundamental property of networks, some components of the biochemical network (and by proxy the encoding genes) tend to interact with one another(Jeong et al. 2000). A genetic interaction (GI) is a functional relationship between two (or more) genes, where simultaneously disrupting the activity of interacting partners results in a phenotype that is significantly different from the phenotype predicted from individual disruptions. The interaction is termed negative or positive, depending on if the phenotype of the GI is stronger or weaker than expectation (Baryshnikova et al. 2013). The search for genetic interactions helps elucidate network structure by identifying parallel routes and serial cascades in the metabolic transformation of substrate to biomass (Costanzo et al. 2010). It illuminates genetic redundancies that sustain biological robustness(Li, Yuan, and Zhang 2010) and reveals synthetically lethal effects when co-inactivation of individually non-essential genes leads to loss of cellular viability(Costanzo et al. 2016). GI analysis helps assign function to genes of unknown function, a category that applies to greater than 30% of coding sequences in the best studied organisms (Gagarinova *et al*. 2016). It has improved predictive models of metabolism to increasingly match *in vivo* behavior (Szappanos et al. 2011; Ma et al. 2018), and has led to the discovery of novel drug targets against debilitating diseases (Farmer et al. 2005) and infectious agents (Côté et al. 2016). Thus, experimental determination of GI holds a foundational role in our understanding of gene and genome function.

Synthetic Genetic Arrays (SGA) have been the long-standing method for genome-scale search for GI and involves robots performing high-throughput mating between two microbial strains harboring single mutations to yield a daughter strain that carries both mutations(Typas et al.

2008; A. H. Y. Tong et al. 2004). Strains carrying each combination of a double gene knockout are individually maintained and assayed for fitness based on colony growth-based image analysis(Takeuchi et al. 2014). Even with extensive automation, this approach consumes extraordinary time, effort, and material resources because fitness assays done even parallelly do not scale well to the reality that the number of GI to be assayed squares with the number of genes in an organism. In recent years, techniques to determine strain fitness based on massively parallel sequencing have gained traction(van Opijnen, Bodi, and Camilli 2009). These methods have been adapted for multiplexed GI analysis by creating transposon mutant libraries in known knockout genotypes(Nambi et al. 2015; DeJesus et al. 2017). While a significant improvement from mating based parallel assays, transposon-insertion-sequencing based GI analysis required maintaining mutant pools for each knockout genotype separately because multiple insertions within a single genome couldn't be associated with each other. Multiplexed and simultaneous screening of yeast GI based on DNA barcode sequencing has now been developed(Jaffe, Sherlock, and Levy 2017) and recently extended to multiplexed protein-protein interaction analysis(Schlecht et al. 2017). This technology requires creating targeted genome-scale gene knockouts of yeast in two strains before performing *en masse* mating of the strains and screening for genetic interactions. Application of the method to a chosen microorganism has several prerequisites - the ability to create precise gene insertions in the genome, transfer of chromosome from one mating strain to another, and precise genomic recombination to link strain specific barcodes. CRISPR/Cas9(Jinek et al. 2012) is a recent milestone in genetic tools and can be implemented in a broad host range(Jiang et al. 2013) to create multiplexed gene knockouts (Y. Tong et al. 2015) or gene expression repression(Peters et al. 2016). However, most bacteria lack non-homologous end joining repair mechanisms(Bowater and Doherty 2006) so nuclease active Cas9 cannot be used to create gene deletions, and effective gene repression is subject to spatial organization of genome and transcriptional activity at a gene locus. Further, combinatorial cloning of two guide RNAs for genome-scale GI probing still requires creation of (N*(N-1))/2 clones by direct DNA transformation into an organism with N genes, which is technically very challenging.

We aspired to create a scalable genetic system for rapidly creating double KOs that could be ported to lesser studied bacterial strains, which would enable us to perform large scale genetic interaction screens in non-model organisms of medical and environmental importance. We reasoned that transposons would be the ideal tool for building such a system because of their broad host range (Wetmore et al. 2015), negligible host factor dependency(Kimura et al. 2016), and minimal components required for functioning(Goryshin et al. 2000; Rubin et al. 1999). For semantic simplification, we consider a functional inactivation caused by a transposon insertion as equivalent to a knockout (KO) although transposon insertions within multidomain proteins act as

caveats to this assumption (Goodall et al. 2018). Inspired by multiplexed transposon insertion mapping methods like Tn-Seq (van Opijnen, Bodi, and Camilli 2009), TraDIS (Langridge et al. 2009) etc., we particularly wanted the new genetic method to be compatible with massively parallel next generation sequencing (NGS) so that double KOs can be exhaustively assayed simultaneously in a cost efficient manner. Based on these design criteria, we have developed 2D Tn-Seq (two dimensional Tn-Seq), a method to assay genetic interactions at the genome scale.

We used two orthogonal transposons to create two independent insertions in the chromosome of a bacterial cell. The insertions can be of any distance from each other, and either transposon inserts into random target locations. It is straightforward to introduce two transposon insertions in the same cell to create a double KO. However, preserving the association between two transposons originating from the same cell is non-trivial when multiple bacterial colonies of random double KOs need to mixed together for NGS library preparation and sequencing. Pooling colonies and lysing the cells for DNA extraction breaks the compartmentalization provided by an intact cell and also shears the contiguity of the released genomic DNA, thus physically separating the two transposon insertions in addition to the mixing with insertion sequences from other colonies. To solve this problem, we innovated a DNA barcode-based solution that preserves the association of transposons arising from the same cell (Figure 1A).

The heart of the 2D Tn-Seq method is the *2D-Tn* transposon, a synthetically assembled linear DNA construct in which a barcoded mariner transposon is nested within a Tn5 transposon (Figure 1B). These two transposon systems are orthogonal, having distinct cognate enzymes and recognition sequences. The barcode (20 bp) is composed of randomly synthesized nucleotides and each molecule of 2D-Tn is expected to bear a unique barcode (one of 4^{20} combinations). First, we create the first dimension of insertional mutants by inserting 2D-Tn into the target bacterial genome as a Tn5 transposon (primary mutants) (Figure 1C). The DNA barcode within each primary mutant is mapped to the transposon - genome junction by NGS. The primary mutants are expanded during cell growth and then, the mariner transposon is induced to jump from within the confines of the primary Tn5 insertion into a new, different location within the same copy of the genome. Because the mariner transposon inserts itself into a random TA dinucleotide in the genome, several thousand unique secondary mutants (second dimension of mutants) are created from each primary mutant (Figure 1C). The DNA barcode travels with the mariner transposon and therefore, insertion locations of secondary mutants can be associated with their primary mutant locations, again by NGS. Positive selection conferred by the independent insertion events of the orthogonal transposons ensures that only valid double KOs are selected. After several refinements in the molecular process of 2D Tn-Seq, we were able to create up to

100 million double KOs (10^4 primary mutants each giving rise to 10^4 secondary mutants) in *E. coli* in a single experiment. We have now demonstrated a successful proof of concept where a pool of 10^4 secondary mutants were linked to their respective parent primary mutations by means of their shared unique barcodes.

Experimental Procedures

Bacterial strains & Molecular cloning

E. coli strain BW38028 (Conway et al. 2014) and its derivative SDB202 (BW38028 Δ*hisD*::loxP) were used in 2D Tn-Seq experiments to create double mutants. SDB202 was constructed by Lambda Red mediated recombination (Datsenko and Wanner 2000). Strains BW25141 (Haldimann and Wanner 2001) and DH10B (New England Biolabs) were used in molecular cloning and for method development. Bacteria were grown on LB plates or LB liquid culture, supplemented with 50 mg/ml Kanamycin, 17 mg/ml Chloramphenicol, 50 mg/ml Carbenicillin (antibiotics from Teknova) as necessary.

Molecular cloning of plasmids was accomplished through GoldenGate cloning (Engler, Kandzia, and Marillonnet 2008), using BsaI restriction enzyme and T4 DNA ligase (New England Biolabs). The 2D-Tn transposon template plasmid p18117 was constructed using the following parts - kanamycin resistance gene cassette from pACYC177 (NEB), Omega interposon from pUT-miniTn5 Sm/Sp (de Lorenzo et al. 1990), λ attP from pAH63 (Haldimann and Wanner 2001), chloramphenicol resistance gene cassette from pACYC184 (NEB), SUMO gene (codon optimized for artificial gene synthesis and is a gift from Hideaki Nakayama, Kyoto Sangyo University). Plasmid pEB001(Brutinel and Gralnick 2012) was internally renamed to pTnMmeI and served as the source of mariner transposase gene. Plasmid pSBFLP (Bowden, Palani, and Libourel 2017) carrying the Flp recombinase under an arabinose inducible promoter was used to initiate FRT recombination in the 2D-Tn transposon.

Tn5 transposon mutagenesis

Linear double stranded DNA flanked by Tn5 mosaic recognition sequences (Tn5 transposon) was prepared by PCR (primers NPP632/NPP633BC) using p18117 as template. The transposon was cleaned using column clean up kits (Zymo DNA Clean & Concentrator 5) to remove any salts or organic contaminants, and then resuspended in molecular biology grade water. Transposon DNA

concentration was adjusted to be between 150 - 200 ng/µl. To 1 µl of transposon DNA, 0.5 µl of Ez-Tn5 transposase enzyme was added and mixed well. The mixture was incubated at 30 °C for at least 2 hours, and 1 µl of the mixture was electroporated into electrocompetent *E. coli* (1.8 kV). Post electroporation, cells were recovered in 1 ml of SOC medium for at least 1 hr at 37 °C before being plated onto LB plates with antibiotics and incubated overnight at 30 °C. Insertion locations of transposon mutants were mapped either by arbitrary PCR (O'Toole and Kolter 1998) or Illumina sequencing.

Creating 2D Tn-Seq double mutants

SB202 (pSBFLP) Tn5 primary mutants were scraped off overnight growth plates and resuspended in LB + 2% w/v glucose. Cells were mixed by vortexing and then subject to the secondary transposition process. Half of the cells were incubated for 1 hr at 30 °C in in LB 2% w/v arabinose to induce expression of FLP recombinase and other half was incubated in LB + 2% w/v glucose as control to repress the arabinose promoter of FLP recombinase. Each sample was split into two, pelleted, and one pellet of each sample was resuspended into M9 minimal media + 2% w/v rhamnose to induce expression of the mariner transposase. The other pellet was resuspended in M9 minimal media + 2% w/v glucose as control to repress the rhamnose promoter of mariner transposase. Cells were washed in the respective media once to remove traces of LB before resuspension. Cell suspension aliquots were incubated for 24 hours at 30 °C for the transposition step on a rotary shaker (250 rpm). During this incubation, there is no cell division because the histidine auxotrophic cells were resuspended in minimal media allowing mariner transposition to occur in viable but non-dividing cells. Any growth during this period will let the mutants that have weak fitness defect to outgrow the mutants with strong fitness defects. The cells were then spun down and resuspended in LB + 2% w/v glucose. An equal volume of 50% v/v glycerol was added, mixed well, and stored at - 80 °C as frozen stock. Cells were later thawed, spun down and resuspended in M9 minimal media + 1% the normal concentration of LB (0.25 gm per liter). The low percentage of LB media was used to limit colony expansion, and thereby reduce the dynamic range of cell numbers in colonies. Cells were then plated on standard petri dishes or glass oven plates on solid media made from the same composition, and supplemented with chloramphenicol and kanamycin. The plates were incubated at 37 °C for 24 hours. The 2D-Tn mutants were scraped off cells into LB + 2% w/v glucose and stored as glycerol frozen stocks.

Promoter trap

The promoter trap transposon was made by PCR amplification and cloning of only the chloramphenicol acetyltransferase ORF and its native ribosome binding site (from plasmid pACYC184, NEB using primers NPP431/NPP433) into the mariner transposon of pTnMmeI (NPP426/NPP427), resulting in plasmid p3CmPT. The trap transposon was introduced into *E. coli* BW25141 as either a plasmid (BW25141 + p3CmPT) or as an electroporated Tn5 transposon (BW25141 - 3CmPT-Tn5) . Colonies from either transformation were selected on kanamycin antibiotic and replica plated on chloramphenicol selection to verify that there was no leaky resistance arising from the trap. The transposase was supplied *in trans* from the plasmid pRham-HisHimarTnp (gift from Hideki Nakayama, Kyoto Sangyo University, Japan) under the tightly regulated rhamnose inducible promoter. Strains bearing the transposon either on plasmid or chromosome and the transposase plasmid were grown to OD_{600} = 0.6 at 30 °C, and 2% rhamnose was added as inducer. The cultures were further grown for 90 minutes and an aliquot of cells was plated on chloramphenicol selection and incubated overnight at 30 °C. Colonies arising with resistance to chloramphenicol is due to transposition of the trap into an actively transcribed region.

NGS library preparation for massively parallel mapping of transposon junctions

Over the course of this project, we followed library preparation protocols from (van Opijnen, Bodi, and Camilli 2009; Lazinski and Camilli 2013; Langridge et al. 2009) for mapping transposon insertion locations. Sequencing of libraries was performed either on Illumina HiSeq 2500 or MiSeq at the University of Minnesota Genomics Center.

NGS Data Analysis

Tn-Seq sequencing data were obtained as fastq files from the University of Minnesota Genomics Center. Sample demultiplexing based on barcodes and adapter trimming were performed using bbduk.sh module of JGI BBTools package (https://sourceforge.net/projects/bbmap/). Alignment to the reference genome was done with Hisat2 (D. Kim, Langmead, and Salzberg 2015). Bioawk (https://github.com/lh3/bioawk) was used to extract transposon insertion positions from SAM output file of Hisat2. Simulations and analyses were performed in Matlab (MA). A Monte Carlo simulation was performed to calculate the standard deviation associated with fitness when expansion factor is varied. For each chosen expansion factor, the mean number of reads per mutant per time point was varied from 100 to 1000 in steps of 50. Reads for 10^5 trials were drawn

from a random normal distribution and the fitnesses calculated (as in (van Opijnen, Bodi, and Camilli 2009)), from which the standard deviation of fitness was derived. Genome statistics for *E. coli* were calculated in Matlab using the NCBI reference sequence in Genbank format.

Results

Electroporation of Tn5 transposome complex creates Tn5 primary mutants in *E. coli*

Delivery of the *2D-Tn* transposon to create uniquely barcoded primary mutants required that there be no *in vivo* expansion of the nested transposon construct prior to insertion into the target organism. Therefore, we sought a transposon delivery system that could insert an exogenous transposon DNA into the bacterial chromosome without the transposon DNA undergoing any form of replication before insertion. We attempted to replicate the successful precedent of using *in vitro* assembled transposome complexes to create random insertions in the target organism's genome (Goryshin et al. 2000; Akerley et al. 1998). Because purified hyperactive Tn5 transposase was commercially available (EZ-Tn5 Epicentre Illumina), we chose the Tn5 transposon as the primary transposon delivery system. The barcoded 2D transposon construct was generated by PCR, with the Tn5 mosaic ends flanking the sequence. This synthetic DNA was incubated with the Ez-Tn5 transposase and the mixture was electroporated into *E. coli*, with transposon mutants recovered by positive selection for resistance to chloramphenicol. These primary mutants were also sensitive to kanamycin, as expected. We confirmed that neither the transposase enzyme nor the transposon DNA could lead to antibiotic resistance when transformed individually. The near-random nature of insertions was confirmed by Illumina sequencing of a library of approximately 10^4 Tn5 mutants (Figure 2). It is estimated that the frequency of multiple insertions in the same cell is at less than 1% (personal communication from Fred Hyde, Epicentre Biotechnologies).

Positive selection ensures recombination-based release of secondary transposon nested within the primary transposon

The mariner transposon which will create the secondary insertion is nested within the Tn5 transposon that created the primary insertion. Having positive selection for the release of the mariner transposon is essential to ensure that the mariner transposon is completely removed from its initial location, and that any chloramphenicol resistance exhibited by the 2D-Tn double

mutant is a result of re-insertion of the mariner transposon at a different genomic locus. We initially chose the bacteriophage P1 Cre-Lox system (Wierzbicki et al. 1987) as the recombination mechanism and the aminoglycoside phosphotransferase (*aph-I*) gene (originally from transposon Tn903 (Oka, Sugisaki, and Takanami 1981)) conferring resistance to kanamycin as the positive selection scheme (Figure 1C). We later switched to the Flp-FRT recombination system from *Saccharomyces cerevisiae (Senecoff, Rossmeissl, and Cox 1988)* because we observed that expression of Flp enzyme was better controlled compared to Cre, which caused a high level of unintended lox recombination even at baseline leaky gene expression (Buchholz et al. 1996).

We first confirmed that the *aph-I* protein could tolerate the translational fusion of a single FRT site to its N terminus. Next, we constructed a chloramphenicol acetyltransferase (*cat*) gene flanked by two FRT sites that were oriented in the same direction (Figure 3A). We placed this FRT - *cat* - FRT construct between the native promoter and the open reading frame (ORF) of *aph-I*. We performed the initial development and testing of the split *aph-I* construct on a plasmid. An *E. coli* strain bearing this construct was as sensitive to kanamycin as a naive control strain, thus proving that separation of the *aph-I* ORF from its promoter disrupted expression of the gene. When the Flp enzyme was supplied *in trans*, the bacterial strain gained resistance to kanamycin and became sensitive to chloramphenicol. This result indicated that the Flp enzyme performed recombination between the FRT sites, resulting in excision of the *cat* gene and re-joining the *aph-I* ORF to its promoter resulting in its expression. We confirmed this hypothesis by PCR amplification and Sanger sequencing of the modified *aph-I* gene before and after Flp-FRT recombination.

Spurred on by the positive result of gaining kanamycin resistance post recombination, we integrated the construct into *E. coli* chromosome by Tn5 mutagenesis. When we tried to repeat the earlier test of recombination, we found the cells not only did not gain kanamycin resistance but also became chloramphenicol sensitive (Figure 3B & C). More perplexing was the fact that the chromosomally integrated *aph-I* construct had undergone successful recombination between FRT sites, and the resulting nucleotide sequence of the *aph-I* was identical to the post-recombination sequence observed when the gene was carried on a plasmid. Through molecular cloning, we constructed a sequence that is identical to the post-recombination *aph-I* sequence and integrated it into the bacterial chromosome (by Tn5 transposition), which resulted in gain of kanamycin resistance. Based on these results, we were encountering a case of gene repression that was heritable, observed only on the chromosome, and required *in vivo* recombination to manifest. While this problem is a very interesting biological mystery, we focused on finding a solution to relieve gene expression rather than investigating the mechanism of repression.

After recombination on the chromosome, the modified *aph-I* ORF is fused to its promoter. The N terminus of the resulting protein is comprised of a FRT site translationally fused to the native *aph-I* protein. We hypothesized that an unknown secondary structure could be formed by the palindromic FRT sequence after *in vivo* recombination and this could occlude accessibility of the 5' end of the gene for transcription or 5' end of mRNA for translation. We further hypothesized that if the 5' end of the gene is made free of such secondary structures, we might be able to recover successful expression of the modified *aph-I* enzyme. We modified the split *aph-I* construct to incorporate the ORF of SUMO protein at the N-terminus of FRT:*aph-I*. The SUMO protein is noted for its favorable expression and solubility characteristics, and is frequently used as a translational fusion to ameliorate heterologous expression of poorly expressed proteins (Malakhov et al. 2004). We also moved the start codon from the 5' end of FRT to the 5' of SUMO, expecting to create a fusion protein that has the SUMO protein at the N-terminus, followed by the translated FRT sequence, and the *aph-I* enzyme at the C-terminus (Figure 3D). We tested the new split SUMO - *aph-I* construct as before, on both a plasmid and integrated into the chromosome. *Trans* supply of Flp enzyme resulted in successful recombination followed by gain of kanamycin resistance and sensitivity to chloramphenicol (Figure 3E). With further optimization of Flp expression control through careful optimization of the ribosome binding site (Bowden, Palani, and Libourel 2017), we had achieved a recombination system that was tightly controlled and could be used as a foundation to build the 2D-Tn system.

In vivo mobilization of mariner transposon leads to secondary transposon mutants

We sought to mobilize the mariner transposon from within the primary insertion into a new random genomic site to create the second gene inactivation. To confirm that the mariner transposon can be mobilized *in vivo,* we designed a promoter trap system. This synthetic transposon, illustrated in Figure 4A, was designed to confer positive chloramphenicol selection when it inserted into or near an active promoter or expressed geneic region on the chromosome. Chloramphenicol resistant colonies were obtained from the strain carrying the trap on a plasmid but not in the strain with the trap inserted in the chromosome.

The creation of chloramphenicol resistant colonies by a plasmid-launched mariner transposon promoter trap provided two key results. First, it was possible to mobilize the mariner transposon from a known location to a random chromosomal location within the same cell, thus validating the

feasibility of *in vivo* secondary mutant creation. Second was the observation that the transposon trap was functional when placed on a plasmid, but not from the chromosome. We hypothesized that because the mariner transposon moves by a cut-and-paste mechanism, it might create a double stranded break (DSB) when it is cut from its source location. It is thought that *E. coli* performs DSB repair almost exclusively through RecA mediated homologous recombination (Wilson, Topper, and Palmbos 2003). Because the recA gene was deleted from the *E. coli* strain used in this experiment, it precludes the use of sister copies of the chromosomes for homology-directed repair (*E. coli* has genome copy number > 1 during exponential growth, thus containing sister chromosomes (Nordström and Dasgupta 2006)). An inability of the strain to repair the transposon induced DSB could lead to lethality. Based on this hypothesis, we chose to focus on modifying the sequence of operations that could lead to successful secondary insertional mutagenesis without damaging the bacterial genome with DSBs.

Our initial idea was to induce secondary transposition followed by recombination at the launch location to remove any additional copies of the transposon from sister chromosomes. Based on the promoter trap results, we inverted the order of operations to first induce the recombination followed by transposition. The recombination will excise a circular DNA molecule from the chromosomal location of the primary insertion without causing lethality to the strain (similar to (Gohl et al. 2011)). The excised minicircle could then act as the source for mariner transposon mobilization. We could readily implement this inverted induction scheme using the existing molecular constructs. Still, we chose to include an origin of replication (*ori*R6K) within the FRT recombination sites so that the excised DNA can replicate as a plasmid in the strain BW25141 that expresses the pir gene. The reasoning behind this inclusion was two-fold. After FRT recombination, the excised plasmid was purified and verified to ensure that the molecule was of the expected size. Second, the replication origin conferred the ability to tune the availability of the transposon source for a period of time suitable for the transposase to perform the transposition. The R6K origin was conditionally replicated in the presence of the *pir* protein (Shafferman and Helinski 1983; Rakowski and Filutowicz 2013) supplied *in trans*, and the suppression of pir expression could be used as a switch to turn off R6K ori replication. However, later optimization of the 2D Tn-Seq molecular construct rendered the transposition process efficient enough to not require the conditional origin of replication (Figure 5).

We validated the scheme of first performing recombination followed by transposition by successfully creating secondary mutants from a library of primary insertion locations. We were able to isolate several secondary mutants resistant to both chloramphenicol and kanamycin and determined the insertion locations. Mapping the insertion junctions by Sanger sequencing

revealed that the mariner insertions were indeed true positives that were inserted in the chromosome. This confirmed that *in vivo* transpositions can be achieved from a chromosomal launch site when preceded by site-specific recombination as outlined in Figure 5.

Secondary transposon mutagenesis from a single genomic location leads to random insertions

We sought to verify that we can create transposon insertions that are randomly distributed across the genome during secondary transposon mutagenesis. Demonstrating the randomness of secondary transposition is foundational to the utility of 2D Tn-Seq because the presence of insertional hot spots or cold spots would prevent genome-wide sampling of secondary insertions, thus potentially failing to reveal significant interactions in the GI analysis. To perform this validation, we used the CRIM site-specific integration system (Haldimann and Wanner 2001) to precisely introduce the 2D-Tn construct in the *E. coli* chromosome at a known location, the intergenically located attB site of λ phage integration (Landy and Ross 1977). The mariner transposon was mobilized from this single genomic location to create transposon insertions mutants throughout the genome. We generated a library of approximately 300,000 mariner transposon mutants and performed Illumina sequencing to locate the transposon insertion positions. While most essential genes did not contain insertions, chromatin structure and DNA-protein interactions can influence accessibility of a target site for insertion. Therefore, we couldn't confidently confirm that there were no hot or cold spots. However, we were able to locate insertions throughout the genome (Figure 6), thus validating the use of *in vivo* mobilized secondary transposition.

2D Tn-Seq Proof of Concept

Having independently verified that we can create random insertions in the primary & secondary dimensions using Tn5 & mariner transposons respectively, we aimed to demonstrate a small-scale but complete implementation of the 2D Tn-Seq method. We created a barcoded 2D-Tn construct, and introduced it into the *E. coli* genome by means of Tn5 transposition to create approximately 10^4 primary mutants, whose insertion locations were mapped by NGS. Then, we pooled the primary mutants and induced the transposition of the mariner transposon. Double mutants were selected on antibiotic supplemented solid growth media and approximately 10^4 double mutants were pooled for NGS analysis of secondary transposon insertion locations. Three NGS libraries were constructed, sequenced, and analyzed:

1. A mapping library was made from the primary mutant pool to map Tn5 transposon chromosome junctions. For each primary insertion, a unique DNA barcode (Tn5 barcode) contained within the Tn5 transposon was linked to the primary insertion location.
2. A library was made from the primary mutant pool to link the Tn5 barcode to the Mariner barcode.
3. A mapping library was made from the secondary mutant pool to map mariner transposon chromosome junctions. For all secondary insertions arising from the same primary mutant, a single unique DNA barcode (Mariner barcode) contained within the mariner transposon was linked to the secondary transposon insertion location.

The insertion locations of the orthogonal transposons within each double mutant colony were linked together by the association between barcodes of the orthogonal transposons (Figure 7), with a success rate of 7% for the current iteration of the bioinformatic analysis. The low association rate needs to be improved significantly to make 2D Tn-Seq useful. This can be achieved by 1) increasing the depth of sequencing 2) allowing for mismatches in barcode sequences so that sequencing errors are tolerated 3) use long-read sequencing to directly map the mariner barcode to the Tn5 transposon junction and avoid the use of linking step (Tn5 barcode to mariner barcode).

With this proof-of-concept, we were able to confirm that the 2D Tn-Seq method was indeed able to create and link two random transposon insertions in the chromosome of a bacterial cell.

Design parameters for a genome-scale 2D Tn-Seq experiment

We demonstrated that the individual steps of the 2D Tn-Seq technique perform as expected, and that we are able to use two orthogonal transposons to create random double insertions in the *E. coli* bacterial genome. Because 2D Tn-Seq allows us to create double mutants at the genome-scale, we wanted to establish parameters for experiments intended to measure genetic interactions within the 2D transposon mutant library. A typical experiment would involve
1. Creating an arbitrary number of 2D transposon double mutants.
2. Pooling the mutants into a library.
3. Conducting a growth experiment under a specified environmental condition
4. Sampling the library at two or more time points

5. Preparation of Tn-Seq NGS libraries from the collected samples followed by massively parallel sequencing.

We addressed the following questions regarding the design of experiment.

1. How many primary and secondary mutants are required to get genome wide coverage of genetic interactions?

The *E. coli* genome is 4.64 Mb in size, of which 88% has been annotated as protein coding sequences. The remaining 12% is composed of intergenic regions and genes for non-coding RNA. The average size of a gene in *E. coli* is 931 bases. For data redundancy, we prefer to sample each gene at least twice for both primary & secondary insertions. Thus for genome-scale coverage, we need to obtain one insertion approximately every 465 bases necessitating at least 10^4 insertions per dimension, necessitating approximately 100 million 2D Tn-Seq double mutant colonies to be generated (10^4 primary insertions each of which gives rise to 10^4 secondary insertions). The Tn5 transposon does not have a strong target sequence preference (Green et al. 2012) unlike the mariner transposon which almost exclusively inserts into a TA dinucleotide (Akerley et al. 1998; Rubin et al. 1999). Therefore, we can expect 10^4 random primary insertions to be created by electroporating the Tn5 transposome complex. While the mariner transposon insertion within a sequence is dependent on the presence of a TA position, the *E. coli* genome contains 212024 TA sites in total at a distance of one every 14 bases (geometric mean of distances between TA positions). Therefore, we expect that every coding sequence in the genome is theoretically accessible for the mariner transposon to create secondary insertions.

2. What is the influence of expansion factor on the accuracy of calculated mutant fitness and the associated cost of sequencing?

When performing a growth experiment to measure fitness of transposon mutants, we collect cells from an initial time point t1 and a later time point t2. After we perform NGS on these samples, we can calculate the relative abundance and therefore, the fitness of each transposon mutant present in the inoculum library. A key value used in calculating fitness of each mutant is the expansion factor (van Opijnen, Bodi, and Camilli 2009), defined as

$$Expansion\ factor\ = \frac{Number\ of\ cells\ per\ unit\ volume\ from\ t2}{Number\ of\ cells\ per\ unit\ volume\ from\ t1}$$

The relationship between expansion factor, the number of sequencing reads acquired per mutant per timepoint, and the accuracy of the calculated fitness value was explored by Monte Carlo sampling-based simulation. We find from Figure 8 that for a given number of sequencing reads,

expansion factor is inversely proportional to the standard deviation associated with the calculated fitness. While a higher expansion factor is preferable to minimize sequencing cost (i.e. number of reads required), it also increases the dynamic range between the library members.

$$Dynamic\ range\ of\ a\ population\ = \frac{Abundance\ of\ most\ represented\ mutant}{Abundance\ of\ least\ represented\ mutant}$$

The consequence of increased dynamic range is that at an economically feasible sequencing read depth, we would be unable to detect slow growing double mutants that carry strong negative genetic interactions, which are most informative of the underlying genetic network architecture (Babu et al. 2011). A low expansion factor preserves a narrow dynamic range but increases the sequencing read depth required to minimize the error associated with the double mutant fitness. An estimate of sequencing cost for a genome-scale 2D Tn-Seq study is provided based on the Illumina NovaSeq which currently (Q1 2018) has the lowest cost per Gb of sequence data (product information from Illumina Inc).

Sequencing cost =

$$\frac{Number\ of\ mutants\ *\ number\ of\ reads\ per\ mutant\ *\ number\ of\ timepoints\ *\ cost\ of\ flowcell}{Number\ of\ reads\ per\ flowcell}$$

An S4 flowcell for NovaSeq produces an output of 8 - 10 billion single end reads at a read length of 150 bp, costing approximately $50,000. To calculate fitness for subsequent GI analysis, we need to collect data from 2 timepoints of the growth experiment involving a library of 100 million mutants. Assuming an expansion factor of 5 (which is neither too high nor too low, for reasons stated previously), we require at least 300 reads per mutant per timepoint to limit the standard deviation associated with calculated fitness under 5%. Using these values, we arrive at a sequencing-only cost of between $375,000 to $300,000. While this value is expected to be at least an order of magnitude lower than the prevailing method based on synthetic genetic arrays, 2D Tn-Seq is still a dedicated undertaking that requires careful planning.

Discussion

The maturity of massively parallel genome sequencing technology over the last decade has led to an ever-expanding universe of genes and their encoded biochemical functions available for our purview (Shendure et al. 2017). Generating high-quality finished microbial genomes from even unculturable and exotic ‘microbial dark matter’ has now become a routine pursuit (Hug et al. 2016; Mukherjee et al. 2017). This sea of genetic information however sorely contrasts against

the sparsity of functional genomic data that we have managed to muster for model microbes. With the invention of 2D Tn-Seq, we expect to accelerate acquisition of GI data.

2D Tn-Seq offers several advantages over the state of the art. It is designed to be portable to other bacteria of importance, and every molecular component (transposon, recombinase, antibiotic selection) can be customized to organism specific versions. This enables genetic interaction studies in clinically and environmentally relevant microbes. The method accelerates the comprehensive screening of genetic interaction libraries. In our target *E. coli* lab strain, we could create 100 million double mutants within weeks whereas the only published comprehensive GI search effort to date has spanned nearly a decade (Costanzo et al. 2016). 2D Tn-Seq drastically minimizes the labor and resources compared to synthetic genetic arrays. Once the 2D-Tn library has been created, standard GI experiments are similar to a Tn-Seq growth experiment and can be done with the same setup and in the same timeframe. The ability to quickly create comprehensive double mutant libraries also affords disposability of these libraries - multiple environmental conditions can be assayed, thereby uncovering condition-specific interactions.

The molecular underpinnings of 2D Tn-Seq could be modified and expanded to new functional genomics applications. Similar to the Slingshot tool available for mammalian mutagenesis (Kong et al. 2010), the *in vivo* secondary transposition system can be used as a generalized mutagenesis tool to create insertions at high densities. We have implemented a proof of concept for this technique, which we have named Saturation Tn-Seq, by generating approximately 6 million mariner transposon mutants (unpublished) that were mobilized from a single CRIM integration (Haldimann and Wanner 2001) at the λ *attB* site of *E. coli*. The method is expected to saturate most non-essential TA positions with insertions, thereby revealing gene essentiality at very high resolution. We can elevate 2D Tn-Seq from the genetic interactions domain to that of Protein - Protein Interactions (PPI) by installing outward reading bait and prey fusion proteins onto the orthogonal transposons. Insertions into coding sequences will create bait & prey proteins that can be selected for interactions. The 2D Tn-Seq method may also serve as an inspiration to modify the aforementioned Slingshot method into a powerful system of creating double insertions in mammalian cell lines. CRISPRs have taken the vanguard in mammalian GI analysis (Shen et al. 2017a) but transposon insertions could be used as promoter traps that only target interactions between expressed genes, thus generating data dense genetic interaction maps. A corollary experiment would be incorporating inducible outward reading promoters within insertions so that gain-of-function analyses can be performed on cryptic sequences.

2D Tn-Seq was invented with the goal of democratizing and accelerating genome-scale genetic interaction studies so that any microbial genetics lab can adopt this technology to probe the gene interaction network of their preferred model organism, while substantially reducing the resources currently required for such an undertaking. The method is also set to take advantage of the falling costs of NGS (Check Hayden 2014). We trust that our successful proof-of-concept of 2D Tn-Seq will enable its adoption towards expediting a functional understanding of microbial genomes.

Figures

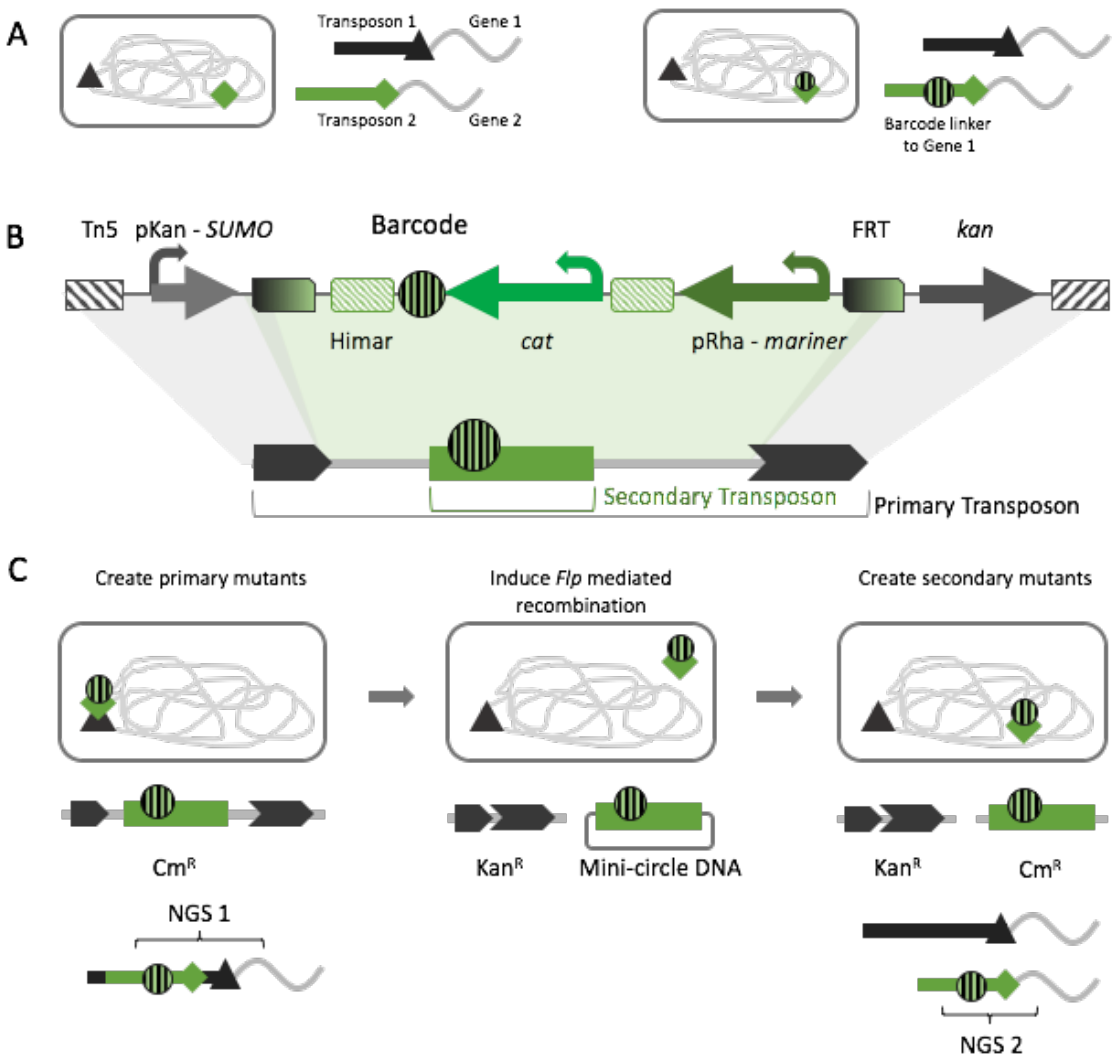


Figure 1: 2D Tn-Seq design and workflow. A) Locations of two independent transposon insertions in a genome can be linked if a barcode encoding the identity of one transposon insertion can be placed within a second transposon insertion. B) Molecular design of the 2D-Tn transposon. Tn5 & Himar indicate the inverted terminal repeats of the Tn5 & mariner transposon respectively. FRT sites oriented in the same direction are recognized by *Flp* recombinase leading to excision of the FRT-enclosed sequence. The *mariner* transposase is expressed from a tightly regulated, rhamnose inducible promoter. *cat* and *kan* are genes conferring resistance to chloramphenicol & kanamycin antibiotics. C) Steps in 2D Tn-Seq. Primary mutants are created,

followed by *Flp* mediated excision of the enclosed secondary transposon. Induction of *mariner* transposase leads to creation of secondary insertions. The DNA barcode moves from the primary insertion location to secondary insertion location. The two locations are linked by the shared barcode. NGS1 & NGS2 are transposon insertion sequencing NGS libraries that provide the barcode - transposon junction information.

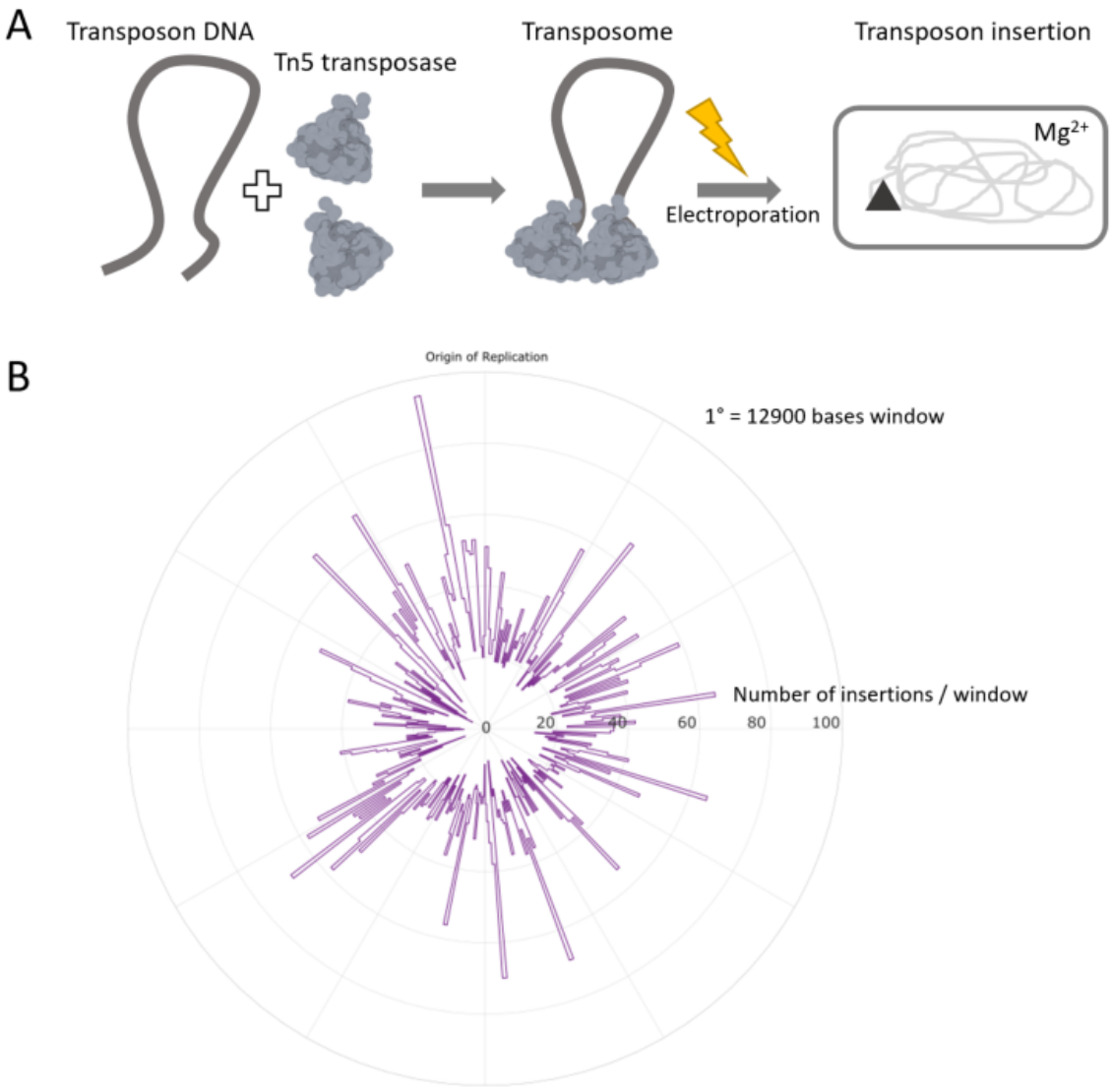


Figure 2: Tn5 transposition for creating primary mutants A) The Tn5 transposon consists of blunt-ended linear DNA that has 19 bp mosaic inverted terminal repeats (ITR) at its ends. The transposon is incubated with purified hyperactive Tn5 transposase (Ez-Tn5, Epicentre Biotech) *in vitro* without any magnesium ions present. Two molecules of the transposase bind to the mosaic ITR to form a stable transposome complex. When this complex is electroporated into bacterial cells, the transposase binds to Mg^{2+} available *in vivo* and inserts the transposon DNA into the bacterial DNA. After insertion, the transposase is dislodged and the nicks at the insertion site are repaired by the host DNA repair system. Insertion of the Tn5 transposon results in a 9 bp duplication of the insertion site on either side of the transposon. B) Polar plot of *E. coli* chromosome showing number of Tn5 insertions grouped into 360 bins, each of size 12900 bp. A

library of approximately 10^4 Tn5 mutants was generated in *E. coli* by transposome electroporation, and the insertion sites mapped by transposon insertion sequencing. The plot shows that Tn5 insertions occur throughout the *E. coli* genome.

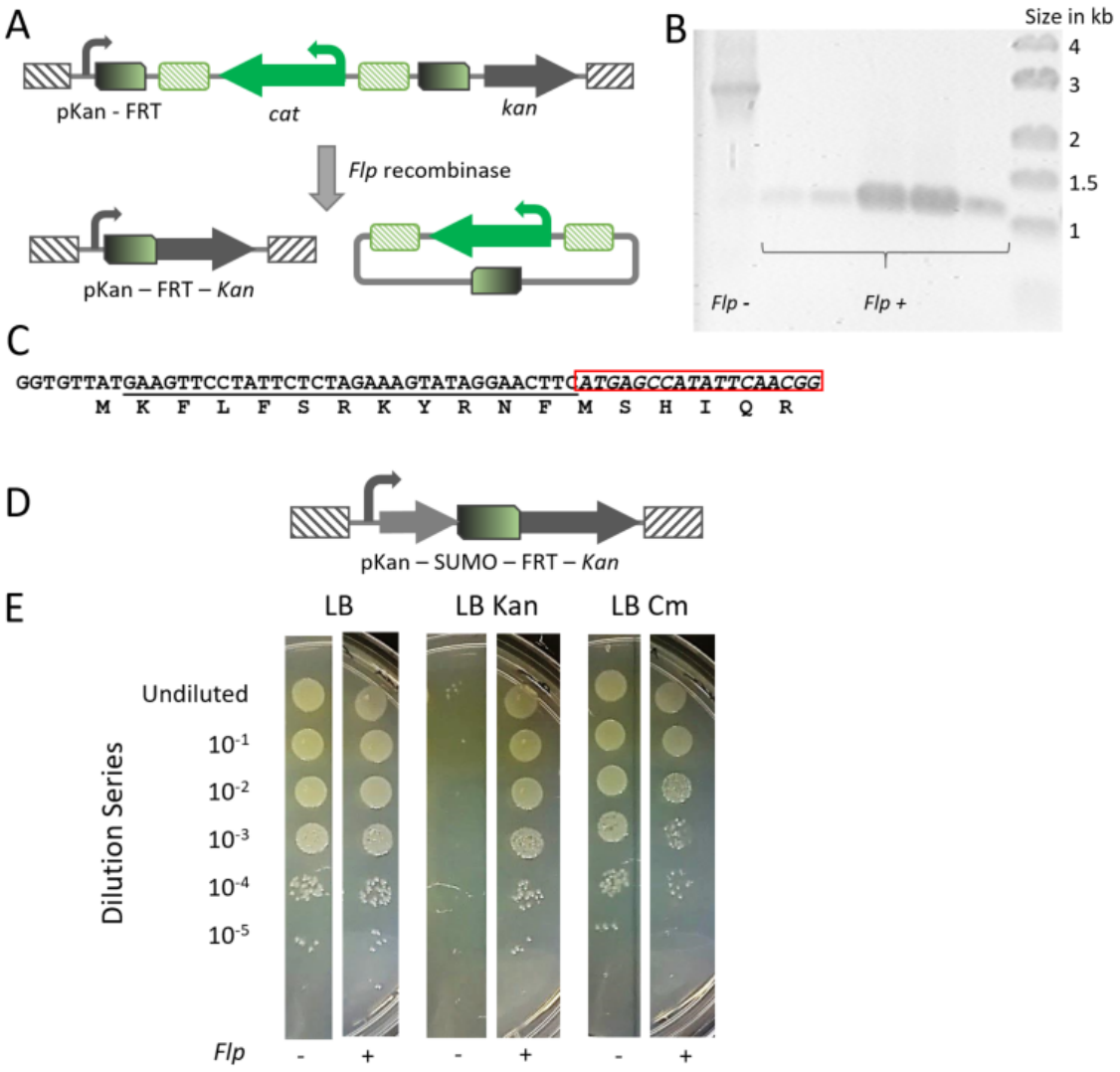


Figure 3: Positive selection for recombination and mariner transposon release A) Design of the molecular construct that was expected to confer kanamycin resistance post *Flp* recombination. *Flp* acts on FRT sites oriented in the same direction, resulting in excision of a mini-circle containing the mariner transposon and leaving an *aph-I* fusion gene at the primary insertion site. B) Agarose gel analysis of PCR product from primary insertion site of Tn5 mutants after induction of *Flp* recombinase. Amplicons were generated by amplification between the pKan promoter and the *Kan* ORF. Lane 1 is PCR amplicon from control strain in which *Flp* was not induced. Lanes 2 - 6 are PCR amplicons from individual primary mutants post recombination. No kanamycin resistance was observed even with successful recombination. Colonies for this test were recovered in LB media without any selection and replica plated for chloramphenicol sensitivity to ensure that colonies had undergone recombination. C) Expected translational fusion

of the FRT-*aph-I* gene fusion. FRT is underlined and the coding sequence of *aph-I* is highlighted in red. Sequence of this post-recombination fusion gene was verified by Sanger sequencing of colonies tested in fig 3B. D) Post-recombination kanamycin resistance was recovered by introducing the SUMO protein at the N-terminus of the FRT-*aph-I* fusion protein, moving the translational start to SUMO. E) Bacterial spot tests show that Kanamycin resistance arising from FRT recombination on the chromosome is tightly regulated and is gained by almost 100% of the cells with *Flp* induction, along with a noticeable decrease in chloramphenicol resistance post recombination.

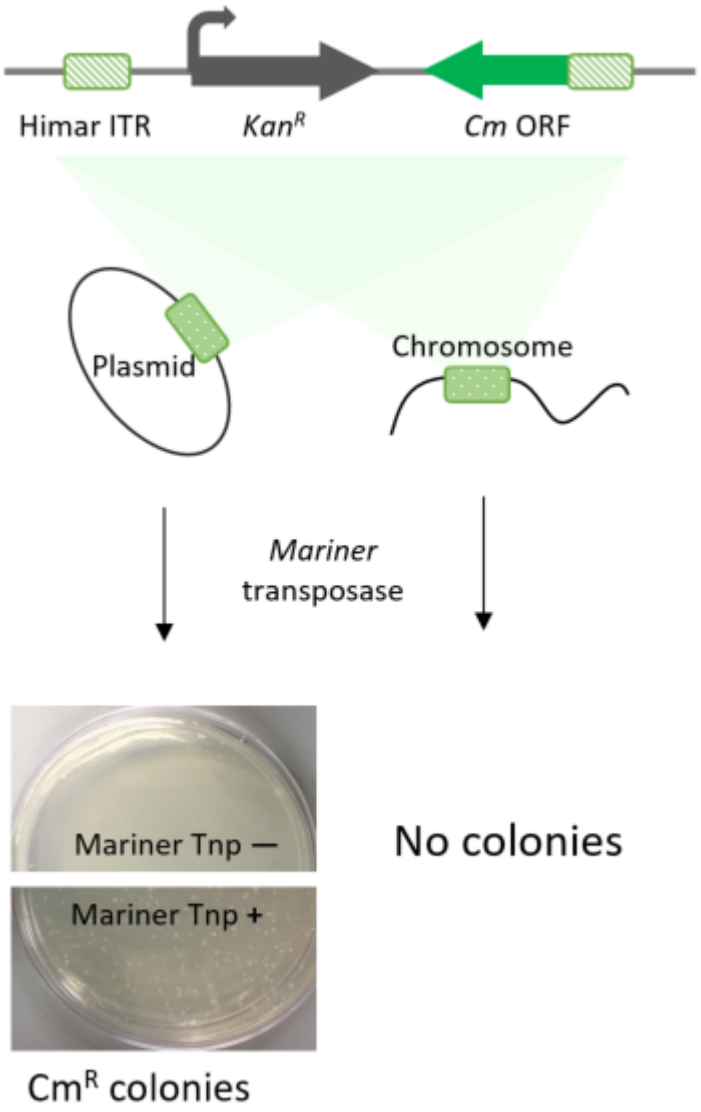


Figure 4: Promoter trap system to test *in vivo* mariner transposition. A) The trap has an inward oriented, promoter-less chloramphenicol acetyltransferase ORF contained within the mariner transposon. Chloramphenicol resistance is gained when the trap inserts into or near an expressed gene after the mariner transposase is supplied *in trans*. Trap transposition was observed only when launched from a plasmid and not from a genomic insertion.

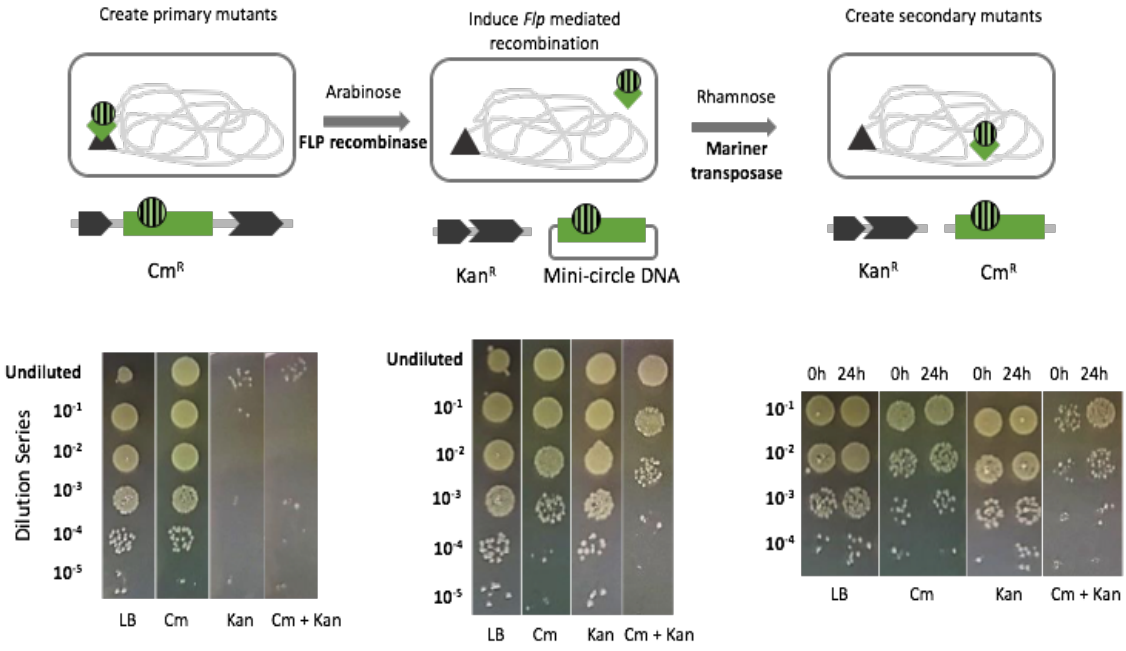


Figure 5: Bacterial growth assays for individual steps of 2D Tn-Seq. Serially diluted cell suspensions were spotted on to media plates. Primary mutants (Tn5 transposon insertions) were resistant to chloramphenicol and sensitive to kanamycin. Induction of *Flp* mediated recombination results in cells gaining kanamycin resistance. 24 hours after induction of mariner transposase, a significant number of cells resistant to both antibiotics (double mutants) are observed. We estimated that ~ 10% of cells that were subject to Flp recombinase and mariner transposase inductions were converted into double mutants.

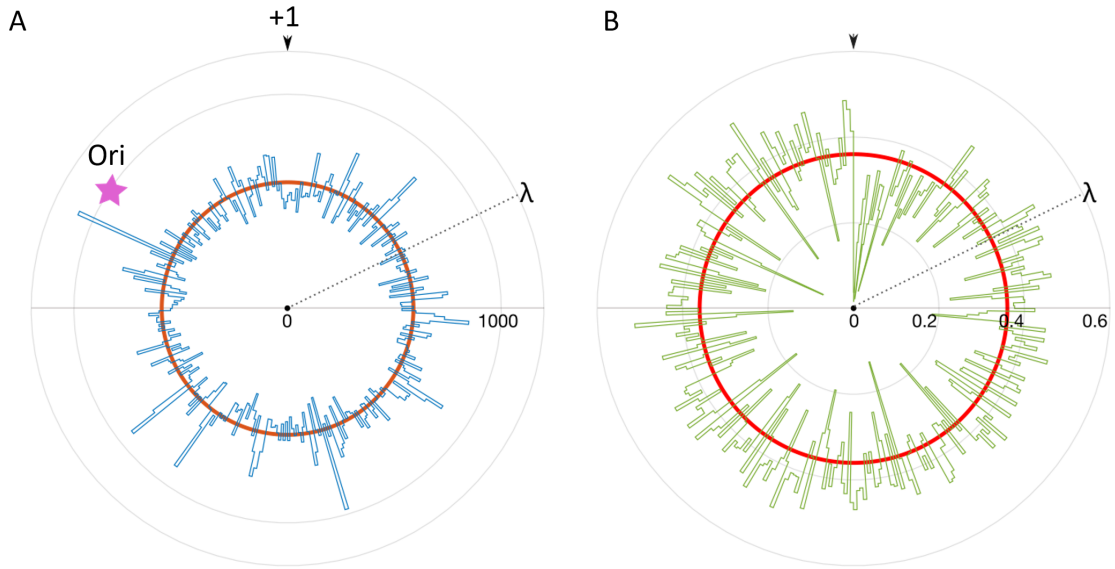


Figure 6: Fraction of TA dinucleotide positions accessed by *in vivo* mariner mobilization. A) Distribution of TA sites in the E. coli genome. TA positions were identified from the MG1655 reference genome and histogram binned into 360 bins, each of size 12900 bases. Average number of TA sites per bin is 588 as indicated by the red circle. B) Fraction of TA sites occupied by mariner transposon insertion in a library of approx. 3 x 10^5 mutants. The red circle is the average fraction occupied by mariner insertions throughout the genome (value = 0.36) for this library. The arrow at the top is the first base of the genome as indicated in the reference genome (NCBI U00096). λ indicates the position of the lambda phage attB integration site. *Ori* is the origin of replication. Bin boundaries are identical for both plots.

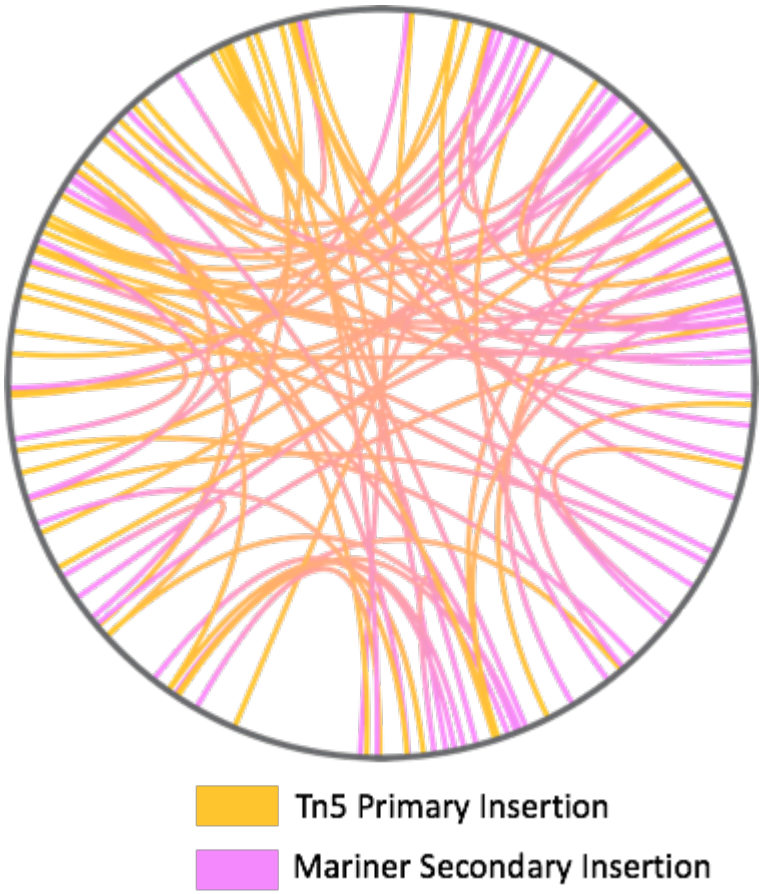


Figure 7: Association of secondary daughter insertions to primary parent insertions. Approximately 10^4 2D Tn-Seq mutants were analyzed to map the mariner insertion locations to the Tn5 locations. 727 associations could be mapped between primary and secondary insertions, of which 77 of the associations are shown in this Circos plot (Krzywinski et al. 2009) as arcs connecting genome positions, with the originating color representing Tn5 insertions and ending color representing mariner insertions. The 12 o'clock position on the plot is the +1 position of the reference genome.

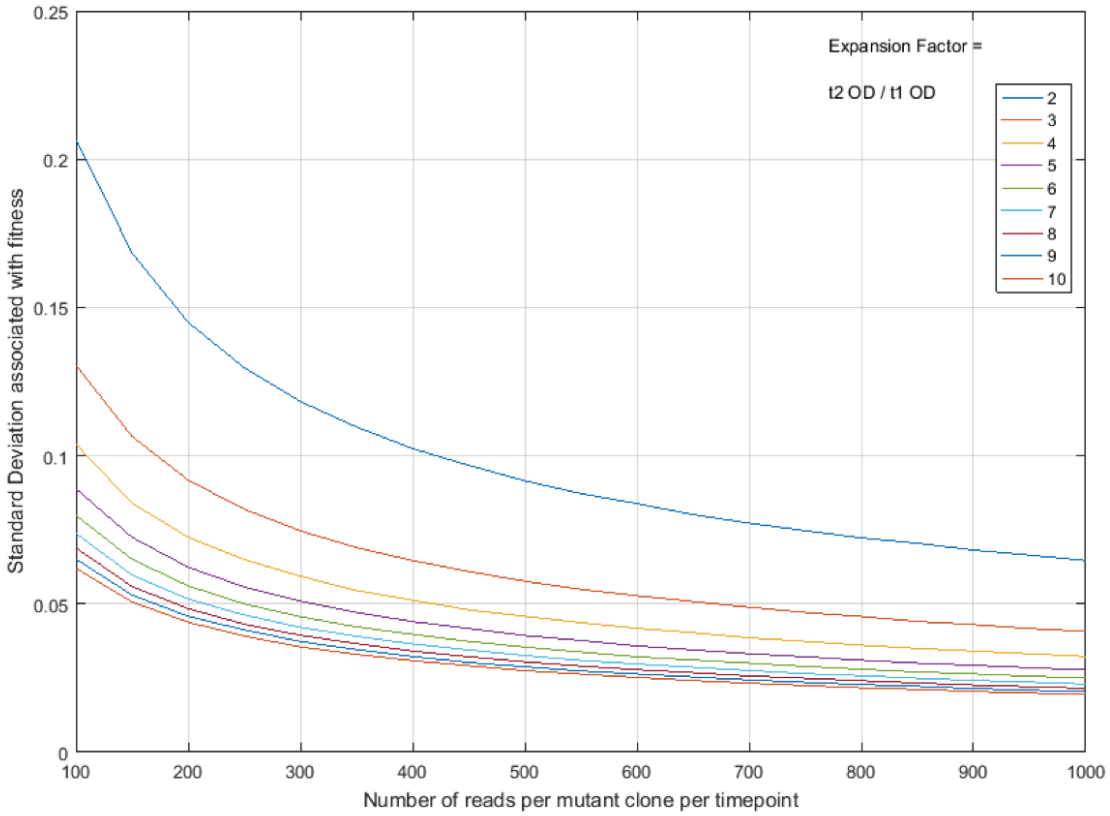


Figure 8: Effect of expansion factor and NGS read depth on the standard deviation associated with mutant fitness. The accuracy of fitness value calculated for a mutant is affected by the expansion factor (length of growth experiment) and sequencing read depth (primary cost of 2D Tn-Seq). For a given number of reads (sequencing budget), lower expansion leads to higher uncertainty in the measured fitness value whereas higher expansion leads to dropout of slow growing mutants.

Tables

Oligonucleotide name	Sequence
SDBMN360 (to introduce 2D-Tn barcode)	TTTTTTGGTCTCNGCTTTNNNNGCAGACCGGGGTCTTA TCATCCAACCTGTTA TGGCAGAAATGACGGGAATTAG
SDBMN214 (to introduce 2D-Tn barcode)	TTTTTTGGTCTCNAAGCTNNNNNNNNNNNNNNNNNNN GGTTATGCAGCGGAAAAGGA
NPP632	/5Phos/ CTGTCTCTTATACACATCT GAAGATGCGTGATCTGATCCTTCAACTC
NPP633BC	/5Phos/ CTGTCTCTTATACACATCT NNNNNNNNNN GACACATGGCATGGATGAACTATACAAAGC
NPP426	NNNNGGTCTCN CCAT CCCGCTCAGAAGAACTCGTCAAG
NPP427	NNNNGGTCTCN TCATCCAACCTGTTATGTGGCGCGGTATTATCC
NPP431	NNNNGGTCTCN ATGG CGGTCACACTGCTTCCGGTAG
NPP433	NNNNGGTCTCN ATGATAAGTCCCCGGTCT CAGGAGCTAAGGAAGCTAAAATGGAG

Table 1: List of oligonucleotides used in this study

Chapter 3

Synthetic Proteins for Peptide based ^{13}C Metabolic Flux Analysis

Nagendra Palani, Steve Bowden, Igor Libourel

Introduction

Metabolic flux analysis (MFA) estimates the rates of reactions or fluxes within a biological system to measure a temporally precise phenotype. Fluxes are the most sensitive reporters of cellular environment and its response to external perturbation, and are held constant at metabolic steady state. MFA is often performed with ^{13}C isotope labeling of input substrate and measuring the resulting distribution of labeled carbon in cellular metabolites at steady state. These measurements can be obtained using gas chromatography - mass spectrometry (GC-MS) from proteinogenic amino acids (Dauner and Sauer 2000), and by nuclear magnetic resonance (NMR) from carbon-based cellular metabolites (Teixeira et al. 2008) . Atom transitions from labeled substrate to different metabolites are iteratively simulated on a core metabolic network, with the fluxes of the metabolic network adjusted in each iteration as an optimization problem (Antoniewicz, Kelleher, and Stephanopoulos 2007). The set of simulated fluxes that reproduce the experimental labeling patterns are then considered to the fluxes that should have existed during the experimental condition of the biological system under study. ^{13}C MFA thus reveals the kinetics of a metabolic network at steady state, thereby providing a deterministic framework for predicting metabolic phenotype using tools like flux balance analysis (Chen et al. 2011).

Conventional analysis of amino acid labeling requires total protein extraction and hydrolysis to yield free amino acids that are derivatized before being injected into the GC-MS. However, hydrolysis of a protein results in loss of protein primary sequence identity, which can be used to identify the spatial location of the protein if that protein has a spatially identifiable expression signature. Thus, flux maps of organelles or individual members of a microbial community cannot be performed in a straightforward manner. In genetically pliable organisms, ^{13}C MFA can be simplified by expressing an innocuous recombinant protein, purifying this protein, and analysing the constituent amino acids to create a flux map which is comparable to that created from total protein hydrolysis (Shaikh et al. 2008). This method can be extended to estimate compartment-

specific flux maps by targeting an array of recombinant genes to different organelles and performing purifications to individually isolate each recombinant protein for GC-MS analysis. Still, current ^{13}C MFA methods are not amenable for multiplexing i.e. simultaneous analysis of several thousand samples similar to genomics methods like RNA-seq (Islam et al. 2011) or Tn-Seq (van Opijnen, Bodi, and Camilli 2009).

Explosive growth of massively parallel sequencing has resulted in a smorgasbord of functional genomics tools. For a growing number of microorganisms and mammalian cells, several genome-scale libraries are available that enable multiplexed assessments of phenotypes. Examples for these libraries are targeted (Baba et al. 2006) or transposon mutants (Jacobs et al. 2003), gene overexpression clones (Kitagawa et al. 2005), and fluorescent protein fusions to native genes for imaging & high-throughput cell biology (Taniguchi et al. 2010). The CRISPR/Cas9 system has been established as generalized genome targeting system, and modules are now available for creating gene knockouts (Sanjana, Shalem, and Zhang 2014), gene activations and repressions (Konermann et al. 2015; Bikard et al. 2013), epigenetic changes (Liu et al. 2016), and targeted mutations in coding sequences (Komor et al. 2016). Multiplexed analysis of several thousand library members is made possible by sequencing the Cas9 guide RNA library (T. Wang et al. 2014) or by sequencing randomized barcodes associated with gRNA (Dixit et al. 2016). Combining ^{13}C MFA with these functional genomics tools will promote a heightened level of understanding cellular function, but current MFA methods that rely on proteolysis are not amenable to multiplexing on a scale comparable to that of genomics tools.

A notable technical advance in experimental flux determination has been the invention of peptide based metabolic flux analysis (Mandy et al. 2014; Ghosh et al. 2014). Here, isotopically labeled proteins are not hydrolyzed into amino acids but digested by a protease into smaller oligopeptides. The peptides are then analyzed in a high-resolution orbitrap mass spectrometer and flux estimates are fitted to the observed peptide mass distributions (Allen et al. 2014). The technology is still maturing, but it is a promising tool that can be applied to a wide range of flux analysis applications like compartmentalized flux maps (example: organelles & microbial communities), cell-cycle resolved flux maps, and performing flux analysis from mass spectrometry imaging.

We wanted to take advantage of the plethora of functional genomics resources available for bacteria by using our experience in peptide-based ^{13}C MFA to create genome-scale libraries that are suitable for simultaneous and cost-effective metabolic phenotyping. To make this concept feasible, we conceived a strategy to tag individual members of the library with unique DNA

barcodes and track these members at the protein level through translated peptide barcodes. The peptide barcodes can also encode flux information if the library is grown to metabolic steady state on an isotope-labeled substrate. Because we have already established methods to fit fluxes to peptide labeling patterns, we should be able to resolve steady state flux maps for each library member that expresses a unique peptide barcode from its DNA barcode.

To realise this method, named FluxSeq, we designed and assembled a plasmid vector for the *Escherichia coli* gene deletion library (Baba et al. 2006) to express a synthetic flux reporter protein. The reporter protein includes a 31 amino acid peptide that is encoded by a semi-random DNA barcode cloned into the vector. The identity of the DNA barcode, the translated peptide barcode and its association to the host genotype was linked by amplicon sequencing of the DNA barcode. The library members can be pooled, grown together on a labeled substrate, and the synthetic reporter proteins purified from the pool. The peptide barcodes within the reporter proteins are then analysed in an orbitrap mass spectrometer to collect the peptide mass distributions that can be used for MFA. We addressed the following challenges to construct a functional FluxSeq vector: an inducible expression system that could function in the presence of glucose; developing a method to select for soluble synthetic proteins; incorporating affinity purification tags that will help us isolate intact proteins. We were able to successfully demonstrate that the FluxSeq vector performs according to its design criteria.

Experimental Procedures

Molecular Cloning

Plasmid cloning was done in the *Escherichia coli* strain BW25141 (Datsenko and Wanner 2000). Transformed strains were selected on LB medium (Difco) supplemented with 50 µg/ml carbenicillin or 50 µg/ml kanamycin (Teknova). Selection of strains expressing the reporter protein was done on 17 µg/ml chloramphenicol (Sigma). Specific Keio collection mutants were grown on LB plates or liquid medium supplemented with 50 µg/ml kanamycin. After transformation with plasmids expressing the reporter protein, Keio strains were selected on plates containing both carbenicillin and chloramphenicol. Liquid and plate cultures of bacteria were grown at 37 ℃ unless specified otherwise.

Plasmid cloning was done using either Gibson cloning (Gibson et al. 2008) or Goldengate cloning (Engler, Kandzia, and Marillonnet 2008). Primers (Table 1) were ordered from IDT DNA (IA) or

Life Technologies (CA). Chloramphenicol acetyltransferase gene from pACYC184 (NEB, MA) (primers NPP500/NPP501) was cloned as a gene fusion to His-SUMO and replaced the Tn5 transposase gene (primers NPP498/499) in the plasmid pRham-HA-Tn5Tnp (gift from Hideaki Nakayama, Kyoto University) to yield plasmid pHSCat. Plasmid pCuminBB (gift from Claudia Schmidt-Dannert, University of Minnesota, St. Paul) was PCR amplified as two fragments (NPP674/675 & NPP676/677) and ligated to remove BsaI sites in the plasmid, resulting in pCuminBB-NoBsaI. The His-SUMO-CAT open reading frame from pHSCat was PCR amplified (NPP686/687) and cloned as a fusion with mCherry gene (primers NPP688/689 and pRham-mCherryBC template) into pCuminBB-NoBsaI (NPP684/685) under the cumene inducible promoter (Choi et al. 2010) to yield the plasmid pHSRFPCat. The Strep Tag II sequence (Schmidt and Skerra 2007) was appended in-frame to the C-terminus of CAT gene by amplifying pHSRFPCat (NPP694/695) and self-ligating to create plasmid pFluxSeq-RFP. To generate a plasmid library encoding peptide barcodes (pFluxSeq - Figure 1), pFluxSeq-RFP was amplified with primers (NPP700/701) and self-ligated to replace the mCherry sequence with the peptide barcode sequence. The peptide barcodes were incorporated as reduced representation semi-degenerate codons in the oligoprimers (Table 2 - designed using (Halweg-Edwards et al. 2016)). The ligation mixture was then transformed into strain BW25113 and Keio mutants by electroporation (Electroporation condition: 1 mm cuvette, 200 Ohm, 25 μF, 1.4 kV on a Bio-Rad GenePulser).

Post transformation, cells were recovered in 1 ml of SOC medium for 1 hour. Cells that were to be selected for expression of functional reporter proteins (library of pFluxSeq) were plated on LB plates that contained 1 mM cumene (Isopropylbenzene / (1-methylethyl)benzene from Sigma Aldrich) in addition to antibiotics.

Protein Expression

Protein expression studies were done with strains BW25141 (pFluxSeq-RFP), BW25113(pFluxSeq-RFP), BW25141(pFluxSeq), and BW25113 (pFluxSeq). An identical protein expression protocol was followed for all strains. Cells from either colonies or frozen stocks were inoculated into a 4 ml LB liquid culture with appropriate antibiotics. After overnight growth, the cell suspension was diluted and inoculated into a flask of 100 ml LB liquid medium to yield a starting OD_{600} of 0.1. The culture medium was supplemented with carbenicillin to keep the plasmids under selection. The flask was shaken at 250 rpm. 100 μl of 1 mM cumene prepared in 100% ethanol was added to the flask when OD_{600} reached 0.3. Cells were grown until OD_{600} of 1.0 was reached. The cells were then harvested and spun down in disposable centrifuge tubes (15 ml or

50 ml) at 5000 g. The supernatant was discarded and the cell pellet was frozen at -20 °C until required.

Protein Purification

The cell pellet was lysed in B-Per Complete Bacterial Protein Extraction reagent (ThermoFisher) following manufacturer's protocol (5 ml of reagent per gram of wet cell pellet). After incubation to lyse the cells, the suspension was centrifuged at 7000 g and the supernatant collected for further analysis. The recombinant protein was purified first using a Ni-NTA resin (Sigma) that binds the His-tag, and then using streptactin resin (IBA Bioscience) that binds to the strep tag following the recommended protocol (https://www.iba-lifesciences.com/tl_files/uploads/bilder/produkte/streptag/Downloads/Manual%20Double-tag.pdf - Pages included in Appendix). Both purifications were done in gravity flow columns at 4 °C on the same day. Whenever required after His-tag purification, the eluted protein was incubated with SUMO Protease (LifeSensors, Inc.) for 30 minutes at 30 °C before proceeding with the Strep-tag purification. After the two purification steps, the eluted protein was concentrated and the elution buffer exchanged to phosphate buffered saline (pH 7.4) using an Amicon Ultra-4 (10 KDa cutoff) centrifugal filter (Millipore).

Whole cell lysates and purified proteins were analysed by polyacrylamide gel electrophoresis (PAGE). Samples were mixed with 0.1% SDS and Laemmli buffer (Bio-Rad), placed in boiling water for 5 minutes, and loaded onto 12% gels (Mini-PROTEAN TGX precast, Bio-Rad). Gels were run in tris-glycine-SDS buffer at 180 V for appropriate time, followed by staining in Coomassie Brilliant Blue R250 (Sigma) for visualization.

pFluxSeq library of Keio knockout strains

Specific strains of the *E. coli* Keio deletion collection (Table 3) were obtained from the Coli Genetics Stock Center (New Haven, CT), and electrocompetent cells prepared for each strain. After transformation with pFluxSeq, colonies were selected on LB plates supplemented with cumene, chloramphenicol, and carbenicillin. Eight colonies were picked for each strain and arrayed onto columns of 96-well plates. The colonies were grown overnight in LB liquid medium supplemented with carbenicillin and kanamycin, and frozen stocks were made for storage at -80 °C. To create a pooled library for protein expression and peptide based ^{13}C MFA studies, freezer stock was thawed and cultures inoculated into fresh 96-well plates containing LB liquid +

antibiotics. The culture was grown overnight and the cells from across the wells were pooled into a single aliquot.
Barcode sequence in the plasmid within each strain was mapped to plate well position by pooling samples similar to (Gohl et al. 2014), and preparing Illumina-sequencing compatible amplicon libraries from the five resulting pools (done by Steve Bowden). Plasmid was purified (Sigma GenElute) from each pool and a two-step PCR protocol was followed to first enrich the barcode sequences and then uniquely index the libraries (FluxSeqBC & Nextera primers in Table 1). The resulting libraries were checked on a Bioanalyzer (Agilent) for size and sequenced on a HiSeq 2500 125 bp paired end run to a read depth of approximately 1 million reads per library. Paired end reads were merged, filtered, and trimmed using BBTools (BBMap - Bushnell B. - sourceforge.net/projects/bbmap/), clustered with CD-HIT (Fu et al. 2012) and the DNA barcodes analysed in Matlab (Mathworks).

Results

Design & functional verification of pFluxSeq plasmid

Primary components of the FluxSeq vector required for expression and purification of the reporter protein were a glucose-independent promoter, an antibiotic resistance protein that will retain its function when fused to another peptide sequence, small purification tags that will facilitate *in vitro* protein purification and also not interfere with the function of the fusion protein, and a peptide barcode sequence that can encode sufficient diversity to be used with practically large library sizes.

The cumene inducible expression system is non-native to *E. coli* and no host factors are required for gene expression repression, allowing for strong expression even in the presence of glucose. This is advantageous compared to widely used sugar inducible expression systems like that of arabinose or rhamnose because the presence of glucose represses metabolism of other sugars (Aidelberg et al. 2014). Also, the small molecule cumene is readily diffusible across the cell membrane and does not require any active import mechanism like that required for sugars, leading to more uniform protein expression across the population. Other small-molecule inducible expression systems like the Tet promoter compare unfavorably to the cumene induction system due to the high cost of inducer (example: Anhydrotetracycline). Testing the cumene induction system confirmed that induction of protein expression was rapid (< 1 hour), and uniform (Figure 2A).

To select functional reporter proteins, the peptide barcodes were fused with a protein conferring antibiotic resistance. This selection scheme ensured that only those reporter proteins that were soluble and functionally active were represented in the FluxSeq libraries. Solubility is a key prerequisite because expressed heterologous proteins could potentially affect cellular function as aggregated inclusion bodies. Further, ensuring solubility enables efficient and uniform extraction of the reporter proteins from the library of cells. Two candidate proteins were considered for the selection scheme - Chloramphenicol acetyltransferase (conferring resistance to chloramphenicol) and beta-lactamase (conferring resistance to β-lactams like ampicillin and carbenicillin). CAT is expressed in the cytoplasm and has been used as a selection mechanism for protein solubility (Maxwell et al. 1999). β-lactamase can tolerate C-terminus protein fusions, and is exported to the periplasm, potentially simplifying the cell lysis step of the protein purification protocol. When expressed under a tightly controlled Tet promoter, CAT displayed very low background expression and was functional with both N- and C- terminii fusions. Therefore, CAT was chosen as the selection system for soluble peptide barcodes. To enhance the solubility of CAT-peptide barcode construct, the SUMO protein solubility tag was fused to the N-terminus of CAT. The SUMO tag is small (~ 110 amino acids), demonstrated to enhance protein folding, and can be scarlessly cleaved by treatment with the SUMO protease (Marblestone et al. 2006). Other solubilization tags like the Maltose Binding Protein (MBP) or Glutathione-S-transferase are much larger in size compared to SUMO protein, and the MBP gene was prone to rearrangements and deletions in attempts to clone it into a high copy plasmid.

The inclusion of protein purification tags was guided by the need for non-intrusive tags that could bind to inexpensive resins. The 6x-His tag (6 amino acids) and the Strep tag (8 amino acids) satisfied the design criteria and yielded highly enriched reporter protein when purified using the recommended resins (Figures 2D & 3).

Optimization of protein expression and purification using pFluxSeq-RFP

The pFluxSeq-RFP vector was used to test and improve protein expression and purification protocols. The striking color of the mCherry protein was used as a visual guide in these experiments. Induction of synthetic reporter protein expression with cumene resulted in cells accumulating mCherry in as little as 2 hours, with the cell culture showing a strong magenta color by 6 hours. When cells were spun down, the color of the cell pellet was uniform (Figure 2A)

indicating that all the cells in the culture were expressing the heterologous protein, unlike the multimodal expression often observed with the sugar inducible promoters that require active transport of the inducer molecule into the cell (Siegele and Hu 1997). After cell lysis and centrifugation, the lysis supernatant retained the magenta color of mCherry indicating that the recombinant protein had been released from the lysed cells. However, the lysis pell*et al*so retained mCherry color, indicating that some cells were either not lysed or there were some recombinant mCherry proteins bound to the cell debris. The lysis supernatant was first run through the Ni-NTA column for the His tag to bind. The magenta color of mCherry was retained on the column resin (Figure 2B) whereas the supernatant flow-through and subsequent washes were all colorless indicating that the recombinant protein was strongly bound to the resin and there was not any leaching during the wash steps. The Ni-NTA column eluate was intensely colored, and the column resin returned to its original color after elution. This confirmed that any recombinant protein bound to the resin was completely eluted. Performing the Strep tag purification with the eluate resulted in similar results - the recombinant protein was bound to the streptactin resin (Figure 2C), while the flow-through and wash solutions were colorless. Elution from the streptactin resin was complete, with the resin returning to its colorless state after elution.

PAGE analysis of fractions (Figure 2D) from each step of the dual purification protocol showed that the most of the cellular protein in the lysis supernatant had been washed away in the His tag purification step. There was almost no background cellular protein that could be detected after the streptactin column wash step. While some degradation of the synthetic reporter protein was observed in PAGE, the discrete bands observed indicated that the synthetic protein was breaking at specific locations. The fragmentation of the full-length protein most probably happened during sample preparation for PAGE because if there was protein fragmentation before or during the purification process, the flow-through and wash solutions would have retained some of the magenta color of mCherry and there would also be protein bands for these samples in the polyacrylamide gel. The results from these experiments with pFluxSeq-RFP show that the synthetic reporter protein can be isolated in high purity and yield for downstream mass spectrometry analysis.

Protein expression from pFluxSeq and purification yields highly pure synthetic reporter protein

Strain BW25113 transformed with a library of pFluxSeq plasmid was used for testing synthetic reporter protein expression in the target genotype. Expression of the reporter protein and

subsequent purification using protocols identical to those followed for pFluxSeq-RFP (Section 3.3) resulted in isolation of recombinant protein at the expected size of approximately 45 KDa (Figure 3). Unlike the mCherry version, the synthetic protein with peptide barcodes did not show any fragmentation during sample preparation for PAGE, suggesting that the large size and the folding requirements of the individual components of the mCherry version might have made it prone to fragmentation at specific locations.

Sequencing of DNA barcodes verifies that peptide barcodes match design criteria

In silico translation of DNA barcode sequences obtained by Illumina amplicon sequencing produced a list of peptide barcode sequences. A sequence logo (Figure 4) generated from the peptide barcodes was used to identify the salient features common to all the barcodes. Synthesis and cloning fidelity of the barcodes was confirmed from the observation that for each position in the barcode, only the amino acids encoded by the designated codon were present. There does not seem to be any major bias in amino acid frequencies at a given position, which is to be expected since the compressed codons are non-redundant, encoding multiple amino acids at equal frequency. Any slight change in amino acid frequency for the same codon at multiple positions would have arisen from the nucleotide incorporation biases inherent in oligonucleotide synthesis.

The peptide barcode was designed to yield 3 fragments of ~ 10 amino acids each when subjected to a tryptic digest. Thus, the terminal amino acid of each fragment was either a lysine or arginine, matching the cleavage preference of trypsin. The constant sequence (WSG) in the middle of the sequence and contained with fragment 2 was formed by the nucleotide sequences used to clone the barcodes. The consistency of this sequence shows that only error-free PCR products were able to successfully ligate during the cloning of pFluxSeq. The near-equal distribution of amino acids at each position validates the assumption that these barcodes can encode a level of diversity exceeding currently feasible methods for evaluating functional genomics libraries.

Discussion

We find it desirable to perform multiplexed peptide based ^{13}C MFA on functional genomics libraries to accelerate our understanding of how genes affect metabolism. To accomplish this

goal, we needed a biopolymer barcoding system that could uniquely be associated with host genotypes and tracked in a multiplexed fashion as both nucleic acids and polypeptides. We constructed a plasmid vector pFluxSeq as the molecular tool to enable multiplexed peptide-based MFA. The vector is functional in *E. coli* and expresses a synthetic flux reporter protein from a small-molecule inducible non-native promoter. pFluxSeq is designed to accept an in-frame DNA barcode library within the flux reporter gene, and expression of the gene results in production of a library of heterologous proteins that contain peptide barcodes translated from the DNA barcode library. The synthetic peptide barcode sequences retain their association to the host genotypes, and act as reporters for amino acid isotope labeling patterns when the cell library is grown on a labeled substrate, thus allowing multiplexed MFA.

We incorporated several components into the design of the flux reporter protein to simplify the process of extracting the protein from cells. We included a solubilization domain and a selection system to retain only those library members that express and fold correctly. We included affinity purification tags at either end of the protein to be used with a double purification protocol that will retain only full-length protein molecules. The orthogonal purification steps also yield the flux reporter protein at high purity with negligible or no native proteome contamination, thus simplifying sample preparation for mass spectrometry. Lastly, the peptide barcodes were designed to yield fragments of a size ideal for orbitrap MS analysis and can report labeling information on all 20 amino acids, thereby contributing to the increased accuracy of any flux maps generated.

While pFluxSeq is designed to be used as a tool for peptide-based MFA, the vector can be ported to other protein expression applications. The peptide barcode sequence can be replaced with a library of protein sequence variants to select for soluble variants. The double purification system will yield only full-length intact variants so sequences prone to targeted endoproteolytic cleavage or post-lysis fragmentation can be excluded. After purification, the accessory polypeptide sequences can be cleaved with targeted proteases like SUMO protease and enterokinase to release the unfettered variant protein. This protein molecule can then be used for functional assays or for crystallographic purposes. The cumene induction system supports regulated protein expression in non-model organisms (Kaczmarczyk, Vorholt, and Francez-Charlot 2013) and the components of the flux reporter protein can function without host dependencies, thereby expanding the utility of pFluxSeq vector to a wide range of bacteria.

Figures

Figure 1: **Map of pFluxSeq plasmid** The synthetic reporter protein has the 6xHis tag at the N-terminus and Strep tag at the C-terminus. Successive purifications using these tags enable recovery of full length protein. Translational fusion of peptide barcodes to chloramphenicol acetyltransferase enables selection for soluble proteins. SUMO fusion enhances solubility and also acts as a site for SUMO protease cleavage. The CymR repressor binds to the Cym operator region to block RNA polymerase transcription, and is released when bound to exogenously added cumene.

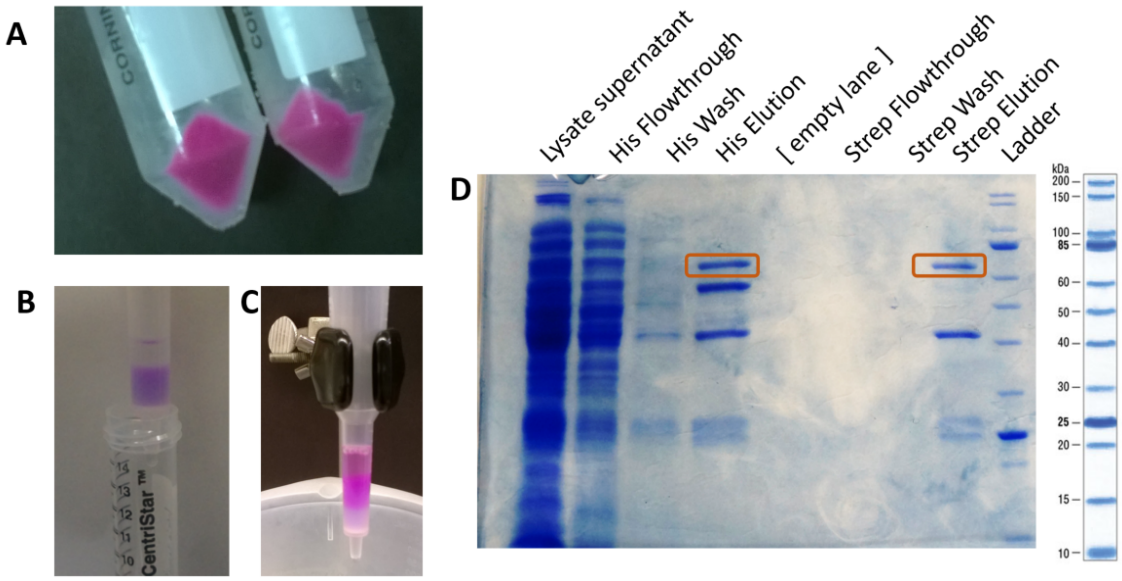


Figure 2: Protein expression and purification of the mCherry fusion protein The vector pFluxSeq-RFP, in which the fluorescent protein mCherry is present instead of the peptide barcode, was transformed into *E. coli* BW25141. This strain was used to visually monitor and optimize the protein expression and purification protocols. The strain was grown in a flask and synthetic reporter protein expression induced by addition of cumene. A) Cell pellets exhibit vivid color from mCherry protein expression after 6 hours of induction. B) Recombinant mCherry fusion protein is bound by the His tag to the Ni-NTA resin. C) Protein eluted from the Ni-NTA column is bound by the Strep tag to the streptactin resin. D) Samples from each step of the dual purification process were analyzed by PAGE. Elutions from the His tag and Strep tag bound columns resulted in highly pure recombinant protein isolate. The boxed bands correspond to the full-length protein (68.2 KDa) and the smaller bands correspond to cleavage products. The mCherry fusion protein was sensitive to the sample preparation steps of PAGE and was subject to some degradation, as indicated by the other bands in the elution product lanes whose sizes add up to that of the full-length protein.

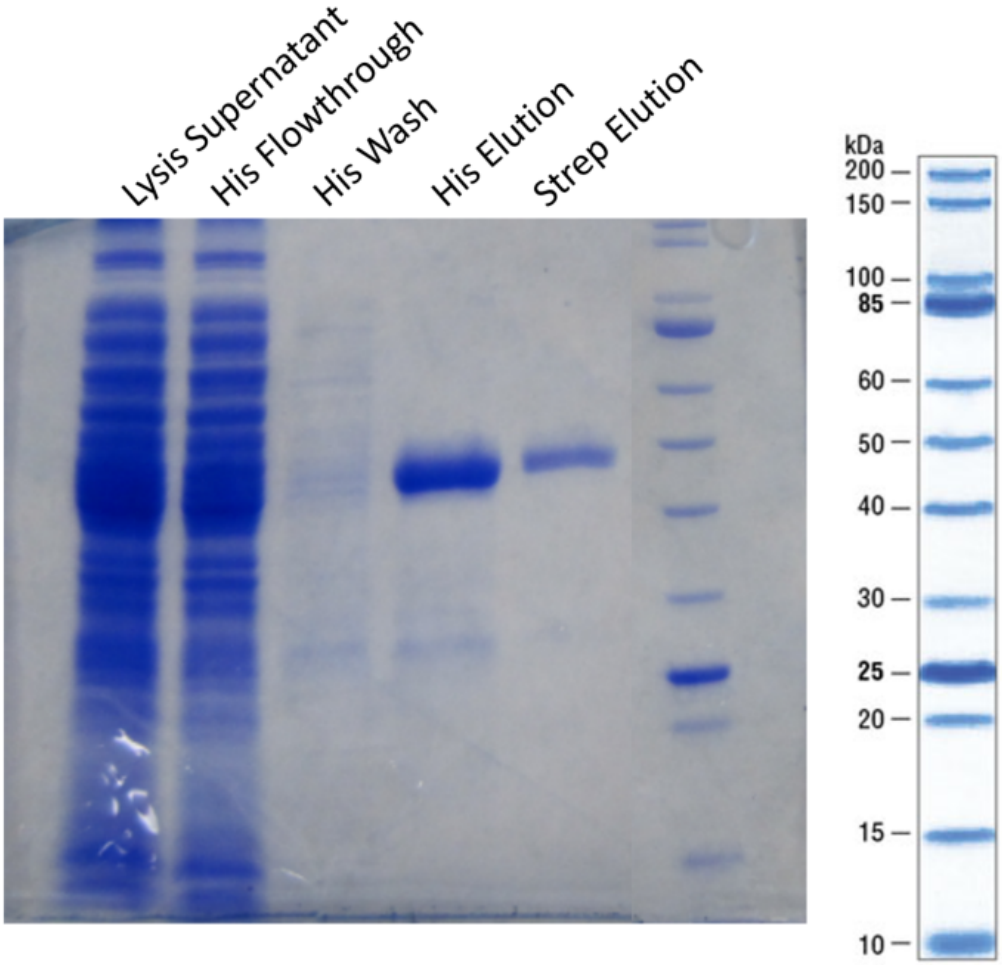


Figure 3: Purification of pFluxSeq synthetic reporter protein with peptide barcode The plasmid pFluxSeq was transformed into BW25113, and approx. 100 colonies were pooled to form a library. The pool was grown in a flask and synthetic reporter protein expression was induced for 6 hours. After dual purification steps, the samples were analyzed by PAGE. For the Strep Elution lane, protein at 1/3rd of the concentration from the His Elution was loaded. The single strong band in the Strep Elution column corresponds to the expected size of the reporter protein containing the peptide barcode (~ 45 KDa), thereby indicating high purity of the protein isolate.

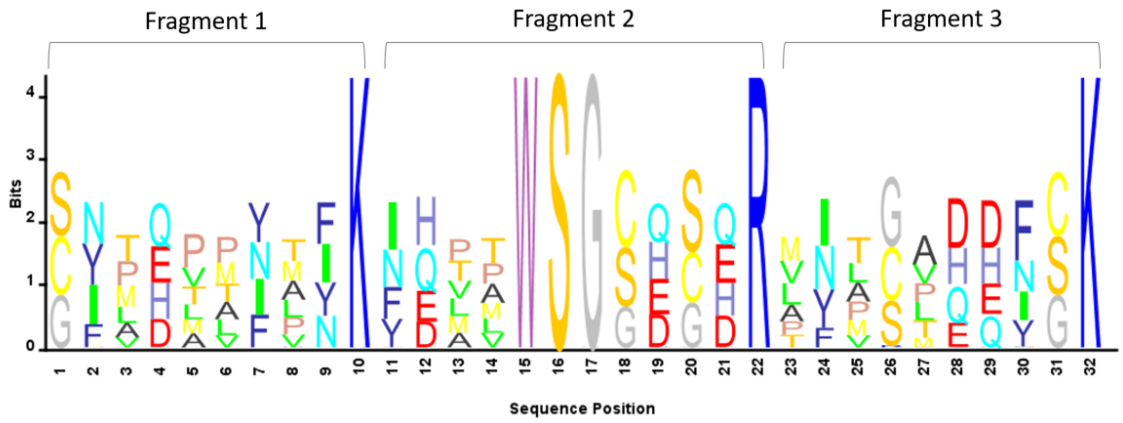


Figure 4: Sequence Logo of peptide barcodes present in the Keio (pFluxSeq) strains DNA barcodes were extracted from Illumina sequencing data and translated to amino acid sequences. The resulting peptide barcodes were aligned and the sequence logo was generated to verify that for each codon position, the appropriate amino acids encoded by the compressed codons were present at similar abundance. Digestion with trypsin should cleave the barcode at lysine(K) and arginine(R) amino acids and yield 3 peptide fragments suited for orbitrap MS analysis. The peptide barcodes incorporated into pFluxSeq encompass a sequence diversity of 2.5×10^{17} variants.

Tables

Oligonucleotide Name	Sequence
NPP498	CATACCACCAATCTGCTCACGATGTG
NPP499	CCACCGCTGAGCAATAACTAGCATA
NPP500	CACATCGTGAGCAGATTGGTGGTATGGAGAAAAAAATCACTGGATATACCACCGTTGATATATC
NPP501	TATGCTAGTTATTGCTCAGCGGTGGGCTTATTATCACTTATTCAGGCGTAGCACC
NPP674	NNNGGTCTCN GGAAGT CGCGGTATCATTGCAGCACTG
NPP675	NNNGGTCTCN TGTTTC CATTGCGCGCTCTGCCTGTGTT
NPP676	NNNGGTCTCN AACA CAGGGCAAGTTGATTGCAGCG
NPP677	NNNGGTCTCN TTCC ACGCTCACCGGCTCCAGATTTA
NPP684	NNNGGTCTCNAAACGATCCCTCCTTCGTTCATAATACAAAC
NPP685	NNNGGTCTCNGGGCGGGGCGTGACCTCGAGGCCCAAGGTTTAAAG
NPP686	NNNGGTCTCNCAAGGAGAAAAAAATCACTGGATATACCACC
NPP687	NNNGGTCTCNCCATACCACCAATCTGCTCACGAT
NPP688	NNNGGTCTCNATGGTGAGCAAGGGCGAGGA
NPP689	NNNGGTCTCNCTTGTACAGCTCGTCCATGCC

NPP694	NNNNGGTCTCN GGTGGCTCCAAGCAGACGCCCCGCCCTGCCACTCATC
NPP695	NNNNGGTCTCN CACCCGCAGTTCGAGAAG TGACCTCGAGGCCCAAGGTTTAAAGC
NPP700	NNNNGGTCTCN CTGA CCA CRB CRB WTS GWW YTT GWW CRB GWW CRB CRB WTS CRB GWW GCH ACCACCAATCTGCTCACGATGT
NPP701	NNNNGGTCTCN TCAGGA DGC SAW DGC SAW CGY VYG WWC VYG DGC VYG SAW SAW WWC DGC AAR GAACAGTACGAACGCGCCGA
FluxSeqBC_Fwd	CGTCGGCAGCGTCAGATGTGTATAAGAGACAG ATCGTGAGCAGATTGGTGGT
FluxSeqBC_Fwd	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG GCGCGTTCGTACTGTTC
Nextera_Fwd	AATGATACGGCGACCACCGAGATCTACAC [i5] TCGTCGGCAGCGTC
Nextera_Rev	CAAGCAGAAGACGGCATACGAGAT [i7] GTCTCGTGGGCTCGG

Table 1: List of oligonucleotides used in this study

Compressed Codon	Exploded Codons	Amino Acid
DGC	AGC	S
	GGC	G
	TGC	C
SAW	CAA	Q
	CAT	H
	GAA	E
	GAT	D
VYG	ACG	T
	ATG	M
	CCG	P
	CTG	L
	GCG	A
	GTG	V
WWC	AAC	N
	ATC	I
	TAC	Y
	TTC	F

Table 2: Expansion of compressed codons Semi-degenerate compressed codons were used in designing the DNA barcodes of pFluxSeq. Each compressed codon encodes several amino acids at equal frequency. Lysine(K) and Arginine(R) act as sites of trypsin cleavage and were not included in the codon scheme. The codon for Tryptophan(W) could not be fit into the compressed codon sets and was included separately in the DNA barcode.

Strain / Gene	Keio Designation	CGSC (Stock Center) Designation	Strain Type	^{13}C flux map available ?
Wild Type	BW25113	7636	Wild Type	Y
rpiA	JW5475	11414	Enzyme KO	N
arcA	JW4364	11117	Trans. Factor KO	Y
crp	JW5702	11596	Trans. Factor KO	Y
fumA	JW1604	9365	Enzyme KO	Y
pta	JW2294	9844	Enzyme KO	N
pfkA	JW3887	10802	Enzyme KO	Y
fruR	JW0078	8378	Trans. Factor KO	Y
zwf	JW1841	9537	Enzyme KO	Y
aceA	JW3975	10858	Enzyme KO	N
sdhC	JW0711	8301	Enzyme KO	Y
ppc	JW3928	10837	Enzyme KO	Y
pgi	JW3985	10867	Enzyme KO	Y
ihfB	JW0895	8917	Trans. Factor KO	Y
tktA	JW5478	11606	Enzyme KO	N
pykF	JW1666	9416	Enzyme KO	Y
mdh	JW3205	10430	Enzyme KO	N
aceB	JW3974	10857	Enzyme KO	N
lpdA	JW0112	8394	Enzyme KO	Y
lrp	JW0872	8901	Trans. Factor KO	Y
sucC	JW0717	8788	Enzyme KO	N
pck	JW3366	10508	Enzyme KO	Y
pdhR	JW0109	8391	Trans. Factor KO	Y

Table 3: List of strains transformed with the vector pFluxSeq Strains with specific gene knock-outs were transformed with the pFluxSeq vector for peptide based metabolic flux analysis. Except the wildtype, strains were knockouts of genes either coding for central metabolism enzymes or transcription factors. Some knockouts have prior published ^{13}C MFA based flux maps.

Chapter 4

Deep Mutational Scanning of Phenotype Arrays

Nagendra Palani*, Daryl Gohl, Archana Deshpande, Kenneth Beckman, Igor Libourel*

* equal contribution

Introduction

Deep mutational scanning (DMS) has harnessed the power of massively parallel sequencing to introduce a systems view of the protein genotype-phenotype relationship (Fowler et al. 2010). Simultaneously assaying several thousand amino-acid variants of a protein has been transformational in our ability to construe how structure determines function (Starita et al. 2013; Jin et al. 2015; Haddox, Dingens, and Bloom 2016; Bandaru et al. 2017). Even with DMS still in its infancy, it promises to revolutionize the functional characterization of poorly characterized proteins. The method has so far relied on phenotypes linked to genotypes to assay protein function through selection. This requirement limits its utility to only those proteins for which a selection scheme is available. Thus, in cases where the measured phenotype is uncoupled from organismal fitness, a new strategy for linking phenotype to genotype is required.

Our work extends the applicability of DMS to proteins that cannot easily be selected for. We have devised a workflow that applies established genomics methods to phenotype screening arrays in order to resolve a protein's structure - function connection. We create a protein variant library by transforming a mutagenised plasmid library into a bacterial expression system. Bacterial colonies are arrayed in a standard format (96 well microtiter plates) and mutants are rapidly screened using available chromometric or chromatographic assays, which enables the method to be readily adapted to existing array-based functional assays. Significant effort has been invested by the enzyme engineering community to develop plate based high-throughput assays that are sensitive and make use of optimized phenotyping workflows. Our method is able to take advantage of such existing assays to provide rich functional information and can generate structure - function mapping even from archival stocks of gene variant libraries.

As a proof of concept, we created an aminoacid variant library of the mCherry fluorescent protein (Shaner et al. 2004) and analyzed how its primary structure affects its fluorescence emission and intensity. A plasmid borne mCherry gene was mutated by error-prone PCR and transformed into an *E. coli* based regulated protein expression system. 9402 bacterial colonies were arrayed in 96-well plate format, and position-linked spectral phenotyping of mutants was obtained using confocal microscopy. While cell sorters have been employed to bin mutants into groups of shared fluorometric phenotype, the number of bins available with current technology is limited and subtle variation in phenotype is not differentiated (Sarkisyan et al. 2016). The genotype of each variant protein in the array was resolved using barcode sequencing of orthogonal sample pools and single molecule real time (SMRT) sequencing to link barcodes to mCherry sequences. We analyzed the mutational scan data using a linear model composed of several protein properties including aminoacid orientation, change in free energy, and aminoacid conservation.

Before the advent of DMS, experimental annotation of functional residues in a mechanistically poorly understood protein was a non-trivial pursuit. Exhaustive characterization of phenotypically plastic residues was almost intractable. Catalytic sites are often assigned based on sequence motif similarities to well-studied proteins of known function, and often require site saturation mutagenesis studies to confirm bioinformatic predictions. Our linear-modeling based approach to analysing mutational scan data of phenotype arrays allows empirical evaluation of functional predictions. Using our approach of linking genotype and phenotype information, we were able to identify the outsized functional contribution of several residues in mCherry, including the chromophore and other previously highlighted residues without any *a priori* information (Shu et al. 2006).

Experimental Procedures

Error-prone mutagenesis of mCherry

The leucine/isoleucine auxotroph *E. coli* strain DH10B (New England Biolabs) was used for cloning and expression of mutagenized mCherry plasmids. Transformed strains were selected on LB medium (Difco) + 50 µg/ml Kanamycin (Teknova). The mCherry open reading frame was PCR amplified (primers mCherry_Fwd / mCherry_Rev1,2, or 3) from plasmid pRSET-B-mCherry (gift from Brett Barney, University of Minnesota) with Taq polymerase (Lucigen), supplemented with three concentrations of $MnCl_2$ to obtain an average of one, two, and three mutations per mCherry sequence. Separately, the transformation vector backbone was amplified from pRham-HA-

Tn5Tnp (gift from Hideaki Nakayama, Kyoto University) using the high-fidelity Q5 polymerase (NEB) and primers pRham_Inv_Fwd + pRham_Inv_Rev (primer sequences provided in table 1). The vector features the rhamnose regulated rhaBAD promoter from *E. coli*, with a pBR322 origin of replication and a Kanamycin antibiotic resistance cassette. During PCR amplification, each amplicon was tagged with a unique barcode consisting of 17 bp degenerate nucleotides in the reverse primer. The PCR products were cleaned with a PCR cleanup kit (Zymo Research) and quantified using a Nanodrop (ThermoFisher Scientific). The vector and barcoded-mCherry insert were Goldengate ligated (Engler, Kandzia, and Marillonnet 2008) for fifteen cycles using BsaI, DpnI restriction enzymes and T4 DNA ligase to obtain the pRham-mCherryBC plasmid library. The library was cleaned up with the PCR cleanup kit and 1 µl of the library was subsequently used for transformation into 25 µl of electro-competent DH10B cells (Electroporation condition: 1 mm cuvette, 200 Ohm, 25 µF, 1.8 kV on a Bio-Rad GenePulser). Transformed cells were recovered in SOC for 1 hour at 37 ˚C before plating on LB + Kanamycin to obtain approximately 100 colonies per standard 100 mm petridish. Plates were incubated at 37 ˚C overnight and stored at 4 ˚C until colonies were picked. Colonies were picked by hand using sterile toothpicks and arrayed into 96 well plates containing LB + Kanamycin. 33 plates were allotted for mutants of each of the three $MnCl_2$ concentrations. The plates were sealed with a Breathe-easy membrane (Sigma) and incubated overnight at 37 ˚C. An equal volume of 50% v/v glycerol was added, mixed, resealed with an aluminum seal and stored at -80 ˚C. In total, 9402 colonies were picked.

Imaging of mCherry strains

Clones for fluorescence imaging were inoculated from frozen stock in 150 µl of M9 medium with 2% w/v rhamnose replacing glucose and supplemented with 50 mg/ml leucine and isoleucine. Plates were sealed with sealing membrane and incubated at 37 ˚C on a plate shaker for 18 h. Membranes were removed prior to imaging to complete mCherry protein maturation by exposure to oxygen. Plates were resealed and fluorescence emission spectra were collected on a Nikon A1 confocal microscope using an automated stage. mCherry was excited at 488 nm and emission spectra were recorded between 500 and 690 nm.

Orthogonal sample pooling and sequencing for spatial tagging of barcodes

An orthogonal sample pooling approach conceptually similar to (Chi et al. 2014) was undertaken to map the unique 17 bp barcode associated with each mCherry strain to its spatial location (Plate – Row – Column coordinates). The strains were pooled by rows, columns, and plates and

plasmids extracted from each pool (Qiagen Plasmid miniprep kit). The 99 plates were processed in two batches of 50 and 49 respectively, resulting in 139 pools in total. The purified plasmid pools were then subjected to two rounds of PCR to prepare the barcodes for Illumina sequencing. In the 1st round, primers ProMut-seq_F / ProMut-seq_R were used to amplify the barcode sequence. In the 2nd round, Illumina index primers (Nextera_Fwd / Nextera_Rev) were used to dual-index PCR products from the 1st round with a pool specific tag. The sequencing libraries were quantified, mixed, and sequenced on an Illumina MiSeq using a 50 bp single-end run. Bioinformatic processing of the sequence data resulted in 87% of the strains being uniquely mapped to a well.

Single molecule sequencing of mCherry clones

The purified plasmids used for barcode mapping were combined into a single sample. The mCherry ORFs including the downstream 17 bp barcodes were PCR amplified (951 bp) and sequenced on a Pacific Biosciences RSII instrument at the Mayo Clinic, Rochester. Subreads from 240-minute movies of two SMRTcells were combined, and the CCS2 (Circular Consensus Sequencing, version 2 – Pacific Biosciences, CA) algorithm was used to obtain a consensus sequence for each productive zero mode waveguide (ZMW) from respective sub-reads. The algorithm was set to require at least 5 sub-reads, 90% output sequence quality and 920 bp length of the output sequence. The output BAM file was imported into Galaxy (https://galaxy.msi.umn.edu) and converted into a multi-fasta sequence file using BAMtools (Barnett et al. 2011). A single consensus sequence was derived from the ZMW consensus sequences that shared the same barcode. The list of barcodes derived from Illumina mapping was searched against the PacBio output file through a local instance of BLAST+ (blastn) (Camacho et al. 2009). The output contained a mapping of PacBio subject entries that matched against the Illumina barcode query entries. For each barcode, sequences from at least two independent ZMWs were required for further processing. Multiple sequence alignment (MSA) of sequences sharing a barcode was performed on a local instance of CLUSTALW (Larkin et al. 2007). The reference mCherry sequence was included within each cluster when performing the MSA. Each aligned cluster was imported into Matlab, and a custom weighted plurality algorithm was used to call a consensus nucleotide at each position, with the reference sequence being assigned half the weight compared to a PacBio sequence to ensure that the output consensus could be called at each position.

Data visualization and statistical analysis

Data handling and analysis was performed with custom Matlab (Mathworks, MA) scripts. Regression analysis was performed on log transformed fluorescent intensities, following visual inspection showing that fluorescence of wild-type clones was close to lognormal (but failed the Lilliefors test). A subset of clones <1% had no detectable fluorescence, which prompted the use of both ordinary and robust linear regression to confirm that observations held both with and without the inclusion of outliers.

Emission spectrum quantification

For calibration purposes, fluorescent emission spectra were collected from (i) wells filled with media (ii) a dilution series of DH10B cells, and (iii) a dilution series of DH10B cells transformed with wild type mCherry. After characterization of media fluorescence, spectra from untransformed DH10B cells were characterized and corrected for media fluorescence using nonlinear fitting. The same process was repeated to characterize mCherry fluorescence whilst correcting for media and cellular fluorescence. The three spectra thus determined were used to fit each of the error-prone PCR amplified clones, thereby determining both backgrounds and relative fluorescence intensity (RFI) simultaneously. Note that unlike taking the maximum fluorescence at a given wavelength, fitting the entire spectrum is fairly insensitive to small spectral shifts or changes in spectral kurtosis.

Results

The mCherry fluorescent protein was previously evolved using directed evolution under strong mutagenic conditions and stringent selection (Shaner et al. 2004; Campbell et al. 2002; Baird, Zacharias, and Tsien 2000; Bevis and Glick 2002). The recent history of very strong selection therefore provides us with an interesting opportunity to investigate genotypic plasticity and functional robustness of a protein after aggressive directed evolution. To investigate the genetic underpinning of mCherry's function, we generated 9402 mutants by transforming *E. coli* DH10B with a plasmid containing a barcoded copy of an error-prone PCR amplified mCherry gene. Individual clones were arrayed in 96 well plates, and an emission spectrum of each clone was collected using an excitation wavelength of 488 nm. In parallel, a pool of all clones was PacBio-sequenced to collect high-fidelity full-length genotypes of the phenotyped wells. To connect the phenotypes to genotypes, wells were pooled by plate, row, and column, and DNA was extracted from each pool. Using barcode sequencing, the barcodes existing in each of the pools were

determined. Barcodes that uniquely existed in a single combination of plate, row, and column pools were used to map a genotype to a phenotype (Figure 1).

Genetic coverage of the mutant screen

The majority (69%) of the clones encoded full length protein sequences and could be uniquely spatially addressed post-sequencing (Table 2). 29% of the clones were single nonsynonymous mutants, 14% contained at least two point mutations, and 4% of the wild type AA sequences contained at least one synonymous mutation. The genotypes were consistent with the previously observed bias of nucleotide transitions in error prone PCR (Lin-Goerke, Robbins, and Burczak 1997). The resulting amino acid mutations covered 20% of all possible AA transitions. Note that most AA transitions require more than a single nucleotide change and were therefore not accessible through error-prone PCR. For only two positions no transitions were found and for 95% of the AA positions, at least three unique AA transitions were observed.

Single amino acid substitutions rarely change emission spectrum

The emission spectra collected for all mutants revealed that very few spectra deviated from wild-type. Deviation from wild-type was detected using the residuum of the scaled spectrum to the wild-type spectrum. The cutoff for detection was manually tuned to minimize false positive inclusion of clones with wild-type amino acid sequences. Using this approach, 71 spectra were identified as altered (Supplementary Figure 1), 49 of which were sequenced with a false discovery rate of 6%. Visual inspection showed that no spectra showed a leading shoulder shifted to a longer wavelength. Two mutants showed mild spectral broadening to longer wavelengths, the vast majority of the altered spectra showed mild shifts to shorter wavelengths, and seven spectra showed an identical strong shift to a shorter wavelength (Figure 2). Inspection of the amino acid sequences of this last group revealed that they all had the M66T transition in common. Residue 66 is part of the chromophore, and its influence on the emission spectrum was characterized previously (Shu et al. 2006). The two spectra broadened to a longer wavelength also shared the same genetic cause (F14Y), indicating that both distinct phenotypes were caused by a single amino acid transition. The mild phenotypic shifts towards shorter wavelengths were caused by a variety of transitions that were observed in altered spectra at least twice including in known (V195M, Q213A) and unknown (F65A, V73I, H75F, L83K, F91A, M150K, V187I, L199I) effector sites. However, changes were small and intensities of all mutants were reduced. Interestingly,

two mutants were found containing two of the above transitions. Both mutants (34K:91L:213L and 65A:73I) showed an additive spectral shift, but 34K:91L:213L (Figure 3A, B) also had twice the intensity of 213L single mutants, suggesting that the 91L mutation stabilizes the 213L transition since 34K is a transition of an external residue with no associated phenotype (Figure 3C, D). Overall, we observed spectral shift in only 11 out of the 236 amino acid positions other than those forming the chromophore, and most AA positions were limited to a single viable transition.

Internal amino acid mutations diminish fluorescence

To investigate the relative importance of AA residues of the mCherry protein, relative fluorescence intensities (RFI) of single AA mutants were averaged per position (Figure 4A). Scaling the log change in fluorescence from wild-type by accounting for the severity of the AA residue change using the BLOSUM62 substitution matrix reduced the variance by 15%, suggesting that the differential effect of AA substitution can partially be explained by the extent of functional redundancy of AA residues. A false color rendering of the RFIs onto the mCherry protein structure (Figure 4B) revealed that mutations in internal amino acid residues almost always reduced, and often diminished, fluorescence. In contrast, permutations of externally directed amino acids barely affected fluorescence.

Linear model of mCherry specific data reveals regions of significance for protein fluorescence

To investigate which parts of the mCherry protein were most significant for protein function, a linear model comprised of local free energy changes (ddG) (Pires, Ascher, and Blundell 2014), residue orientation (Shu et al. 2006), BLOSUM62 amino acid substitution (Henikoff and Henikoff 1992), and residue positions important for surface interactions (Wall, Socolich, and Ranganathan 2000) was developed. Orientation of amino acids explained 33% of the observed intensity variance whereas ddG or BLOSUM62 amino acid substitution accounted for 12% and 15% respectively. Visual inspection of a false-color rendering of the residuals on the mCherry protein structure (Figure 5) revealed that the chromophore and the amino acids that it interacted with constituted the largest stretch of over-predicted RFI, confirming that alteration of active site residues was particularly detrimental to protein function. We further observed that the majority of under-predicted AA residues fell into three classes: (i) residues that were oriented on the surface of the protein, (ii) residues that were in tight turns of a β-sheet. Over-predicted RFI correctly identified individual AAs and protein regions of significance that were previously reported (Shu et

al. 2006). For proteins that have not been studied in detail, such information may be of great help in finding catalytic centers or regulatory domains.

Discussion

Using deep mutational scanning for arrayed protein libraries, we have uncovered several novel residues that are not near the active site of the mCherry protein and might influence protein function structurally, if not biochemically. In contrast to genetic characterization of pre-selected phenotypes, parallel characterization reveals protein regions of little importance in addition to amino acid residues of outsized significance. For instance, although only a small set of amino acids was known to result in spectral changes of red fluorescent proteins (Shaner et al. 2004), our analysis demonstrated that this was not because of the modest scope of the AA transitions that were scanned during directed evolution, but a consequence of the rarity of spectral changes resulting from mutations. Such information could be valuable in protein engineering efforts that might then employ targeted mutational strategies for modifying protein function.

Error-prone PCR is a simple method to generate mutants but it does not effectively sample all possible single mutations at every position. However, technical advances in creating defined and saturating single-mutant libraries (Starita et al. 2015; Haller et al. 2016) will enable near-exhaustive probing of the effect of every amino acid substitution for each residue position. Thus, deep mutational scanning of an average protein of length 300 amino acids can achieved by generating a library of approximately 20,000 single mutants (~209 96-well plates) that has a 97% probability of being complete (Firth and Patrick 2005). Our strategy of tagging mutant ORFs with short barcodes allows one to take advantage of more concise sample pooling schemes (Gohl et al. 2014) for large libraries, and recent innovation in single molecule long read sequencing (Goodwin, McPherson, and McCombie 2016) allows obtaining high-quality full length sequences of gene-length libraries without the need for haplotype phasing, an essential step in assembling gene sequences from short read sequencing (Stapleton et al. 2016). Other applications for empirical determination of amino acid functional importance would be in understanding structure-function relationships in *de novo* designed proteins, and in promiscuous enzymes. Deep mutational scanning of phenotype arrays is a flexible and generalized workflow that enables linking protein function to its underlying structure.

Figures

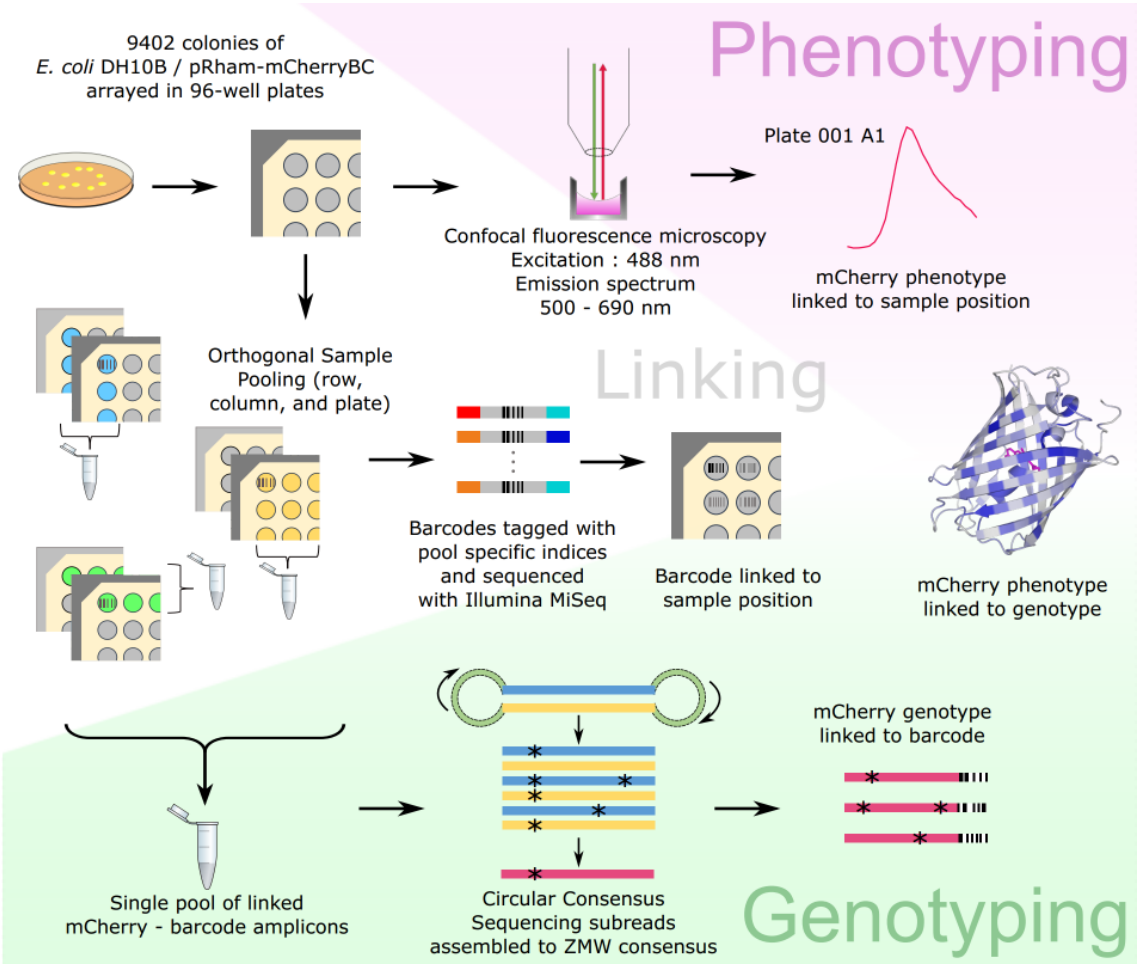


Figure 1: Workflow to link fluorescent protein structure to function. mCherry gene amplified by error-prone PCR was cloned into a barcoded expression vector backbone and transformed into *E. coli* DH10B. Position linked phenotyping was performed by measuring emission spectra of plate-arrayed colonies excited by 488 nm laser on an automated stage confocal microscope. In the parallel linking step, colonies were pooled in an orthogonal scheme, and the barcodes from each pool was amplified as an Illumina sequencing library. Pool libraries were sequenced on a MiSeq, and the barcodes were resolved to their plate positions. Genotyping was performed concurrently by single molecule sequencing of the contiguous mCherry – Barcode fragment library derived from a single pool of all colonies. Barcodes associated genotype to position, and position linked phenotype to genotype, thus resolving a structure – function map for mCherry.

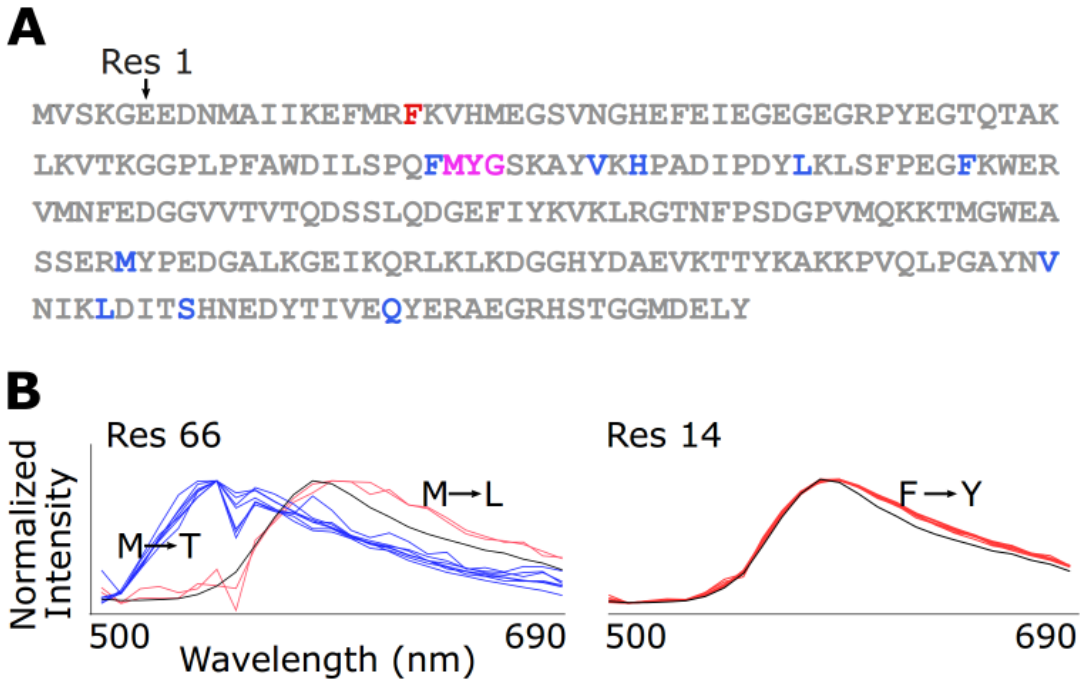


Figure 2: Identification of mCherry residue positions that influence fluorescence emission spectrum Structure – function mapping for mCherry identified residue positions which when mutated cause significant spectral shifts. Several residues marked in blue predominantly lead to blue-shifted spectra when mutated whereas very few residues (in red) lead to red-shifts. Mutating the chromophore (magenta) can either increase or decrease the peak emission wavelength depending on the amino acid transition. (B) Examples for amino acid transitions at select residue positions leading to spectral shifts.

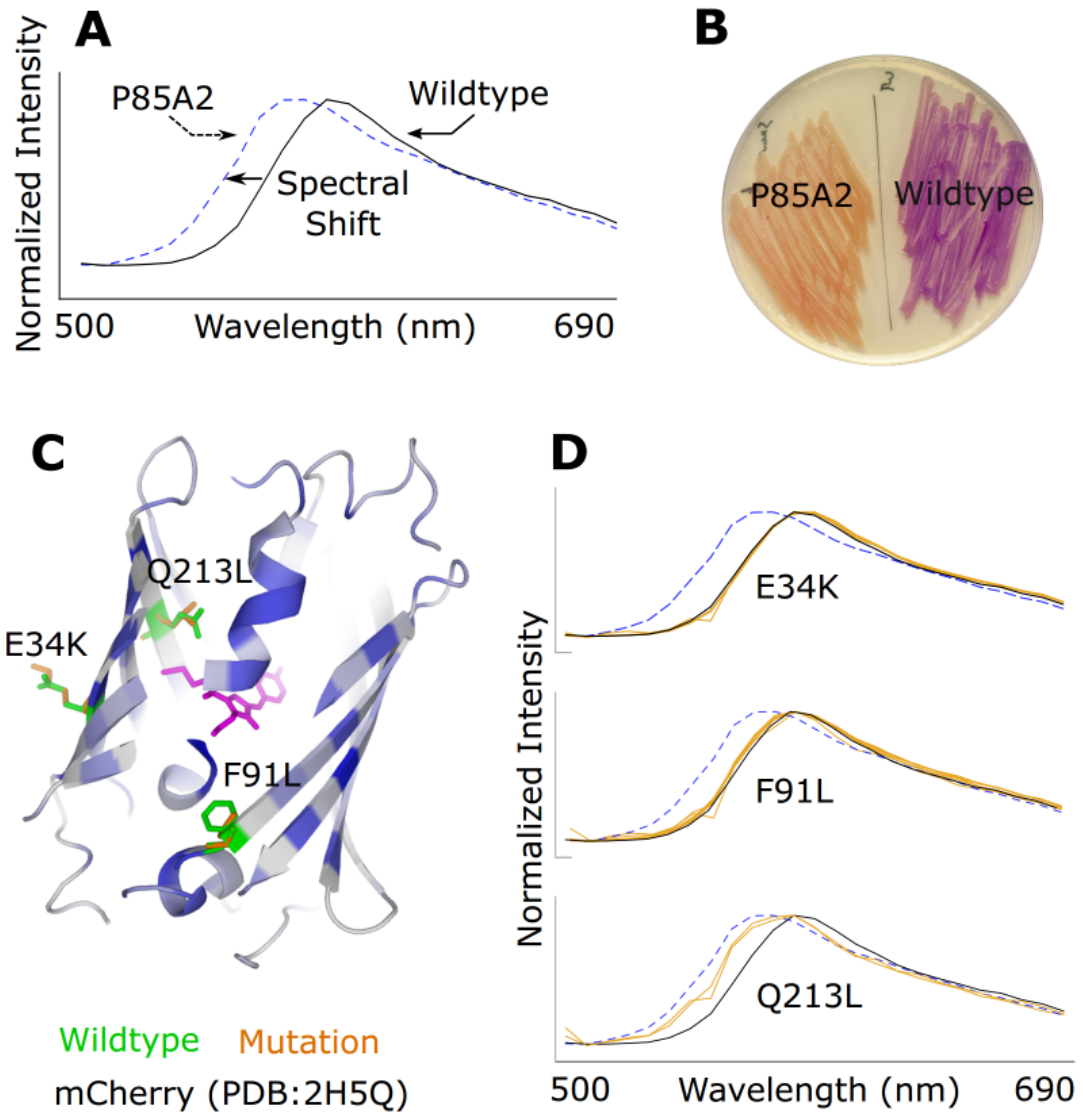


Figure 3: Additive contribution of individual mutations to a phenotype shift From the phenotype data, we identified a mutant (P85A2) whose peak emission was blue-shifted while retaining intensity similar to wildtype. (A) Normalized emission spectrum is compared to the best-fit wildtype spectrum. (B) Strain P85A2 was plated from frozen stock and the strain phenotype was clearly blue-shifted compared to wildtype. (C) Genotype of P85A2 includes three mutations – F91L & Q213L with sidechains interacting with the hydrophobic core of the β-barrel, and mutation E34K with sidechain exposed to the environment. (D) Individual emission spectra of the constituent mutants were compared to the spectra of P85A2 and wildtype. E34K does not seem

to have a role in effecting phenotype change whereas Q213L & F91L induce smaller phenotype shifts that would need to be additive to achieve the phenotype of P85A2.

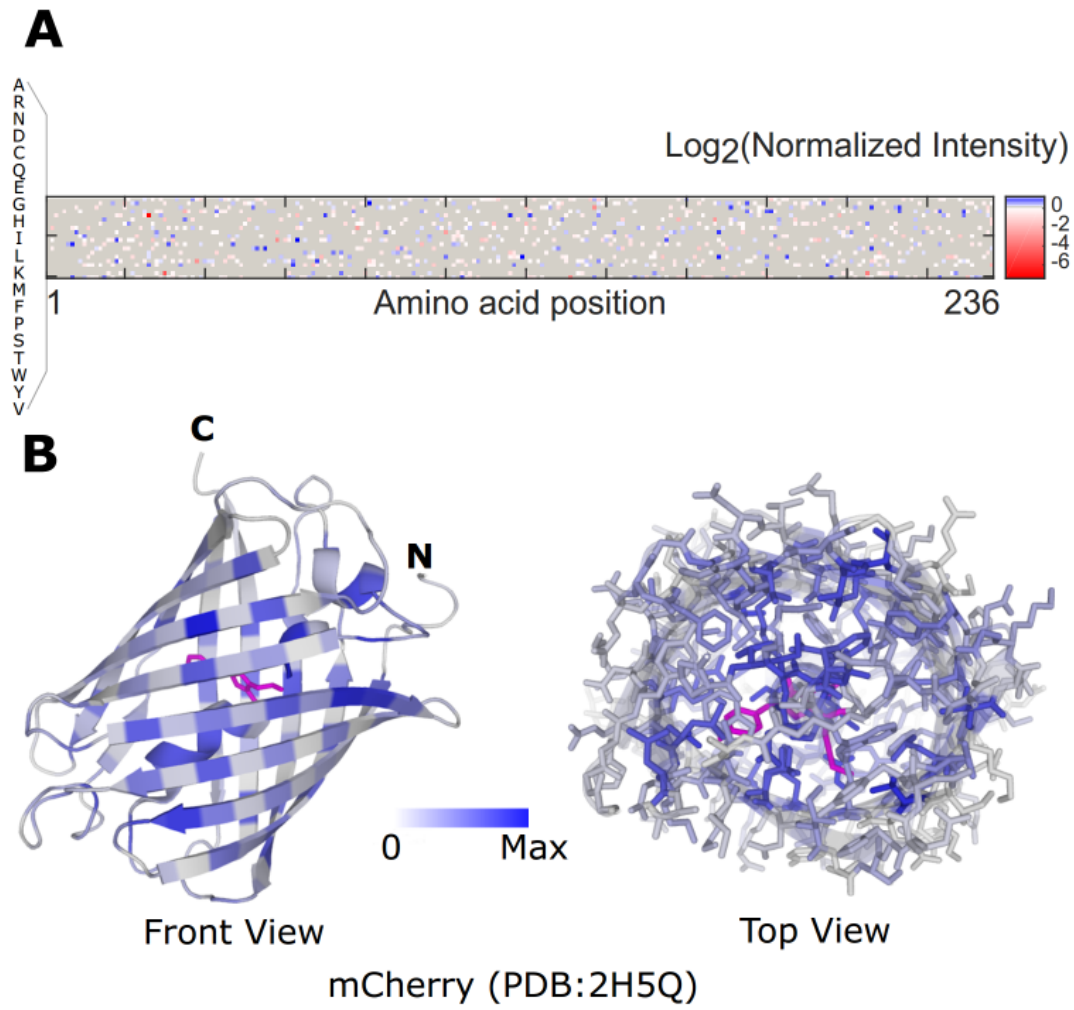


Figure 4: Phenotype severity of mutations in mCherry (A) Heatmap showing the fold change in fluorescent intensity corresponding to each observed amino acid transition indicates that most amino acid transitions negligibly affect emission intensity. (B) Mean fold change in fluorescence intensity per position was plotted on the mCherry structure (PDB:2H5Q), revealing that residue positions with sidechains internal to β-barrel were highly influential in determining mCherry brightness. Residue positions colored white had no effect on intensity when mutated whereas blue shaded positions had the maximum effect. The chromophore is shaded in magenta. Note that fluorescent intensity was not corrected for variation in protein abundance.

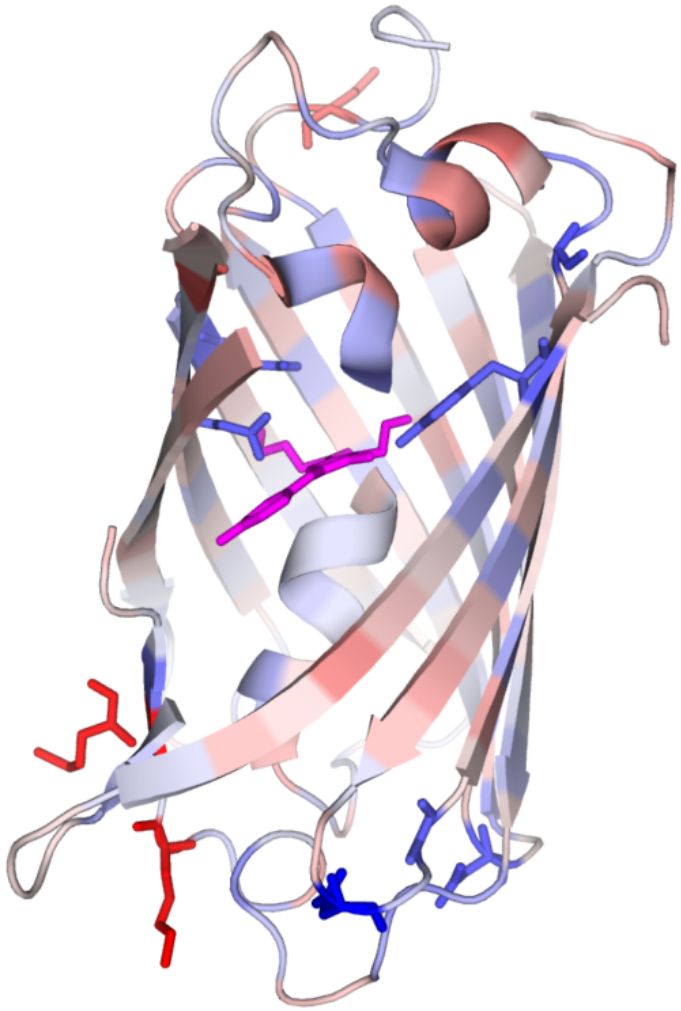
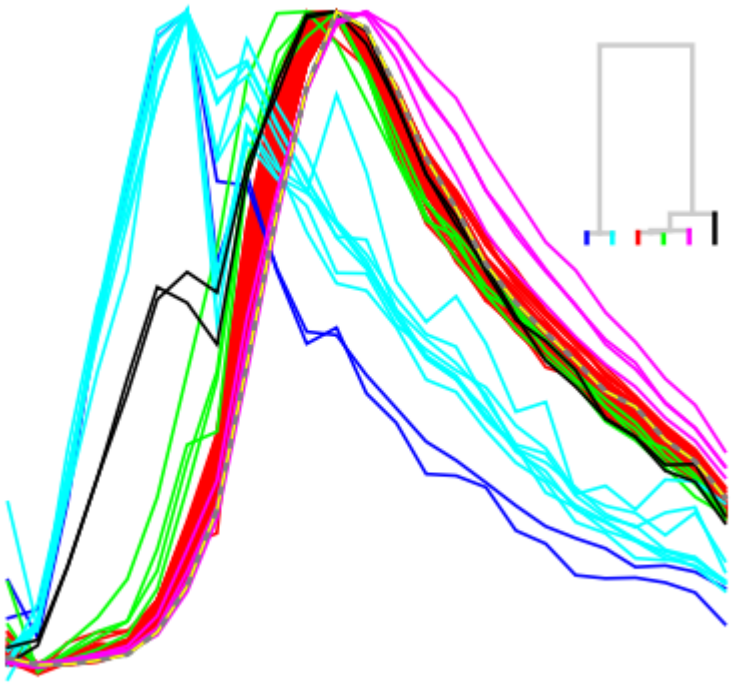


Figure 5: Predicting residue positions that affect phenotype A linear model was constructed to explore the combined contributions of ddG, BLOSUM62, spatial orientation and surface interaction of residues towards predicting phenotype. Residuals from the model were plotted on the mCherry structure in false color, where red is under-predicted and blue is over-predicted relative fluorescent intensity (RFI) respectively. In addition to the chromophore, residues in blue have over-predicted RFI corresponding to current knowledge that these positions are highly influential in determining mCherry phenotype.

Supplementary Figures



S1: **Spectra classification of mutants** Emission spectra of mutants were grouped and overlaid on the wildtype spectrum (dashed line). Most mutants do not exhibit shift in spectrum (red), whereas any shift observed is predominantly towards the lower wavelengths (cyan). Very rarely, red-shifted spectrums occur (magenta).

Tables

Primer Name	Sequence
mCherry_Fwd	NNNNGGTCTCN TATGGTGAGCAAGGGCGAGGAGG
mCherry_Rev1 mCherry_Rev2 mCherry_Rev3	NNNGGTCTCN GTGGT GGAGAAGGTGGCAGCAGCCAACTCAGCTTC NNNGGTCTCN GTGGT TTATAAC GTGGCAGCAGCCAACTCAGCTTC NNNGGTCTCN GTGGT CCACAAT GTGGCAGCAGCCAACTCAGCTTC
pRham_Inv_Fwd	NNNNGGTCTCNCATATGTATATCTCCTTCTTATAGTTAAACAAA ATTATTTCTAGAGG
pRham_Inv_Rev	NNNGGTCTCN CCAC NNN NNN NNN NNN NNN NN CATATG GCAGTTATTGGTGCCCTTAAACG
ProMut-seq_F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGAGTTGG CTGCTGCCAC
ProMut-seq_R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGCACCAA TAACTGCCATATG
Nextera_Fwd	AATGATACGGCGACCACCGAGATCTACAC [i5] TCGTCGGCAGCGTC
Nextera_Rev	CAAGCAGAAGACGGCATACGAGAT [i7] GTCTCGTGGGCTCGG

Table 1 List of oligonucleotides used in this study. All oligonucleotides were ordered from IDT DNA (IA).

Data Processing step	Value	% of colonies picked
Number of colonies picked	9402	100.00
Unambiguous barcodes mapped to wells	8187	87.08
Barcodes after cleanup (duplicates/template)	8150	86.68
Barcodes mapped to PacBio sequences	7767	82.61
Sequences that had >= 2 independent reads	7013	74.59
Sequences with ORFs (minimum 95% length of mCherry)	6638	70.60
Sequences with ORFs exactly matching mCherry size (236aa)	6531	69.46

Table 2 Amount of usable data at each stage of bioinformatic processing of sequencing data.

Chapter 5

Genome-scale CRISPR Interference in *Escherichia coli*

Nagendra Palani, Igor Libourel

Introduction

Transient control of gene expression in a microorganism allows us to probe the dynamic responses of its metabolism. Suppressing or activating transcription of a gene from its baseline expression during the growth period can inform us on how organismal fitness is affected by perturbations to the transcriptome. Until recently, conducting RNA interference or gene repression screens (Y. Ji et al. 2001; Meng et al. 2012) in prokaryotes occupied a niche due to lack of precision tools like shRNA libraries that are available for eukaryotes (Bernards, Brummelkamp, and Beijersbergen 2006). With the development of programmable CRISPR-based methods, microbial systems biology is now poised for growth in the functional genomics arena.

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) form a bacterial immune system and act as a defense against parasitic mobile elements like bacteriophages and broad-range transmissible plasmids (Barrangou et al. 2007). In combination with the Cas9 RNA-guided DNA endonuclease protein, transcribed CRISPR forms a ribonucleoprotein complex that recognizes and cleaves non-self DNA entering a bacterial cell. Immunity against intruders is remembered by capturing a segment of the invading genome sequence and incorporating it as a CRISPR within the host chromosome. Thus, CRISPRs act as records of memory for specific sequences. The first mechanistically explained CRISPR/Cas9 system was from *Streptococcus pyogenes (Jinek et al. 2012, 2014; Sternberg et al. 2014; Anders et al. 2014)*, and was quickly developed into a genome editing tool when it was discovered that any arbitrary DNA sequence containing a Cas9 specific protospacer motif could be targeted for nuclease activity by co-expressing a non-coding CRISPR guide RNA (gRNA) whose sequence matched the target DNA (Cong et al. 2013). Heterologous reconstitution of the SpCas9/CRISPR complex within cells allowed precise gene deletion and creation of knockout strains.

The ease of designing and synthesizing CRISPR guide sequences for Cas9 targeting has led to an explosion of genome engineering applications in mammalian systems. While the native Cas9 protein has endonuclease activity and creates precise double stranded breaks in the genome leading to gene disruptions, the nuclease-deactivated dCas9 protein which only binds to the target sequence without cutting it has been the basis of a plethora of applications. Fusion of specific protein domains to dCas9 has allowed us to target these proteins to a specific sequence, resulting in methods for gene repression (Qi et al. 2013), gene activation (Gilbert et al. 2013), epigenetic modifications (Hilton et al. 2015), base modifications (Komor et al. 2016) etc. The CRISPR gRNA determines the genotype of the cell in which it is present, and can also serve as the source of NGS-based phenotype readout. gRNAs are designed to be unique to avoid targeting multiple sites and thus they serve as unique barcodes that differentiate the genotypes in a pool, making pooled phenotype analysis of CRISPR libraries feasible. Using microarray synthesized oligonucleotide libraries encoding for CRISPR gRNA, (d)Cas9 based screens have been performed at the genome-scale targeting every single gene in several metazoan genomes (T. Wang et al. 2014; Shalem et al. 2014). Further, CRISPRs are naturally present as an array of sequences in bacteria and this property extends to synthetic CRISPR systems, enabling multiplexed and combinatorial targeting in functional genomic screens (Shen et al. 2017b).

While it has become routine to conduct large-scale CRISPR based screens in mammalian cell lines, bacterial versions of such screens are not yet mainstream. CRISPRs have the potential to accelerate functional genomics and systems biology studies in microbes thanks to their ease of design and use. However, applying Cas9 based gene disruption to bacteria is not straightforward because most bacteria do not have a non-homologous end joining mechanism for DNA repair (Shuman and Glickman 2007). Targeting Cas9 to a chromosomal segment causes a double stranded break that can be lethal unless a homologous repair mechanism, native or heterologous, is triggered along with providing a homologous sequence as repair template (Jiang et al. 2013). The practical difficulties of this requirement hinder the use of Cas9 for creating genome-scale gene deletion libraries in bacteria. The dCas9 version however has facilitated multiplexed gene repression (Qi et al. 2013; Peters et al. 2016). Although engineered transposons are widely used to quickly create gene inactivation libraries, they cannot be used to transiently suppress expression. At present, CRISPR based mechanisms are the only feasible way to perform microbial gene repression screens at the genome scale. A defining advantage of CRISPR guided RNA interference (CRISPRi) over gene deletion is that a gene essential for survival can be repressed at any stage of growth, using inducible expression systems, to observe the consequence of depleting the functional molecule (RNA or protein) encoded by that gene. In

addition, the compactness of the components required for CRISPRi targeting make the system portable to most bacteria that can be transformed with an exogenous plasmid (W. Ji et al. 2014).

CRISPR-driven repression of gene expression works by recruiting dCas9 to the target gene sequence, where the bound dCas9 occludes the movement of RNA polymerase thus preventing mRNA synthesis. Several concerns arise regarding the performance of genome-scale CRISPRi screens. If gene repression leads to reduced organismal fitness, then it can impose a strong pressure to select for mutants that have either deactivated the suppression system or introduced compensatory mutations. Replicate experiments done on different days can introduce variability in fitness measurements because of differences introduced by stochasticity in the expression of CRISPRi components. We sought to address these concerns in a CRISPR/dCas9 repression screen in *Escherichia coli*. We wanted to find out from the genome-scale repression screen if we can detect 1) the frequency of escape from repression by a genotype under selection and 2) variability in repression across biological replicates for the same genotype. To accurately measure fitness phenotypes resulting from CRISPRi repression, we constructed a genome-scale gene repression library using a pooled oligolibrary coding for guide RNA targets. To address within-sample variability, each gRNA was combined with gRNA associated lineage barcodes (GLBs) that uniquely tagged independent colonies of the same genotype, thus conferring the ability to record the fitness phenotypes of isogenic lineages in addition to the bulk fitness of that genotype. We then asked if any outliers within the lineages could be identified and removed to reduce the errors associated with fitness measurements. Lastly, we also explored the effectiveness and utility of transient repression compared to constitutive CRISPRi.

Experimental Procedures

E. coli strain DH5a was used for plasmid cloning and genome-scale gene repression experiments. For molecular cloning, bacterial cells were grown on LB plates containing 17 µg/ml Chloramphenicol, 50 µg/ml Kanamycin, or 50 µg/ml Carbenicillin antibiotics (Teknova) as appropriate.

gRNA library design

DNA sequences encoding guide RNA expression were designed computationally using Matlab (Mathworks) and synthesized as an oligo library (CustomArray Inc, WA). The *E. coli* MG1655 genome (NCBI nucleotide sequence U00096) was used as reference to design the CRISPRi

gRNA library. For each coding sequence in the reference genome, a single guide RNA was identified for synthesis as follows. Nucleotide sequence spanning -30 to +150 was extracted from each CDS with the first base of the start codon set to +1. Within this sequence, all subsequences of length 30 and starting with the dinucleotide CC (antisense of the Cas9 protospacer motif NGG) were identified and 15 bases at the 5' end (5' - N_4-N_{19}) compared against the reference genome using BLAST (Camacho et al. 2009). Sequences that had secondary matches of at least 13 out of the 15 base query were discarded. Valid sequences closest to the +1 position of CDS were retained, and further screened to ensure that the potential gRNA oligos did not contain the BsaI restriction site (GGTCTCN^NNNN). The selected gRNA oligos were reverse-complemented to target the non-template strand of the CDS, based on published research demonstrating increased gene repression when compared to targeting the template strand (Qi et al. 2013; Bikard et al. 2013). The oligos were then tagged with primer targeting sites for PCR amplification of the synthesized pool and BsaI restriction sites for cloning the amplified library into a plasmid vector. A control set of synthetic gRNA without any targets in the reference genome was also included in the oligolibrary to serve as the baseline when calculating fitness for gRNA with valid targets. In total, gRNA for 4153 coding sequence targets and 4 synthetic control sequences were synthesized.

Oligolibrary cloning

A two-plasmid system was developed to perform the genome-wide gene repression screen. The plasmid pdCas9-bacteria was a gift from Stanley Qi (Addgene plasmid # 44249). This is a p15 *ori* based plasmid expressing the deactivated *Streptococcus pyogenes* Cas9 nuclease (dCas9) under the Tet-inducible promoter. Plasmid pTra-crRNA is a pBR322 *ori* based plasmid that was the target vector for cloning in the guide RNA library. pTra-crRNA was assembled as follows. DNA sequence encoding constitutively expressed *S. pyogenes* tracrRNA and the constitutive guide RNA expression promoter (from Addgene plasmid #46569) was ordered as a gBlock from IDT DNA (sequence in Appendix). The plasmid pRham-mCherry (this work, Chapter 4) was used as template to amplify the Kanamycin + pBR322 *ori* fragment using primers NPP420/NPP421. The amplified fragment and the gBlock DNA fragment were assembled into pTra-crRNA by Gibson cloning.

To prepare the oligo library for cloning into pTra-crRNA, the single-stranded library was converted into dsDNA and PCR amplified using primers NPP260/NPP261 (PCR conditions: Q5 polymerase, Annealing at 69 °C for 10 s, Extension at 72 °C for 15 s, 10 cycles). The β-lactamase gene fragment from pBAMI-GFP was amplified using primers NPP342/NPP397. A 12 bp degenerate

sequence is included within the primer NPP342 and serves as a DNA barcode (diversity - 4^{12} variations) that can distinguish bacterial colonies that carry the same guide RNA. The gRNA library, the barcoded β-lactamase fragment and the pTra-crRNA plasmid were mixed equimolarly and treated with BsaI restriction enzyme & T4 ligase to perform Goldengate cloning. After the restriction digestion - ligation reaction, the enzymes were heat-inactivated and the reaction product was cleaned by drop dialysis (https://www.neb.com/protocols/2013/09/16/drop-dialysis). The cleaned product (plasmid pCRISPR) was concentrated approximately ten-fold using a centrifugal vacuum concentrator (Speedvac, ThermoFisher Scientific). The gRNA plasmid library was then electroporated into electrocompetent *E. coli* DH5a strain already harboring the pdCas9-bacteria plasmid (100 ng of plasmid library per transformation) (Figure 1A). After 1 hour of post-transformation recovery at 37 °C, cells were plated on M9 minimal media plates containing chloramphenicol & carbenicillin antibiotics to select for both plasmids. In addition, some of the plates were supplemented with 1 μM anhydrotetracycline (aTc) to induce expression of dCas9 to mimic constitutive expression of dCas9. Cells were grown overnight at 37 °C, pooled, and frozen as glycerol stocks at -80 °C. Approximately 50,000 colonies were generated for each growth condition (with and without aTc supplement).

Bioreactor experiments

E. coli libraries containing the CRISPRi plasmids were aerobically grown in bioreactors (New Brunswick Bioflo 110) controlled by custom software (Burdge and Libourel 2014). Libraries from frozen stock were thawed, washed once in M9 medium, re-suspended in M9 medium, and injected into the bioreactor. Enough bacterial cells were inoculated to get an initial OD_{600} of 0.02. Libraries were grown in M9 minimal medium or M9 + 2% Casamino acids medium. Libraries constitutively expressing dCas9 were supplemented with 1 μM aTc throughout the duration of the growth experiment, whereas libraries subjected to timed induction of dCas9 were supplemented with 1 μM aTc when OD_{600} of the culture reached 0.05. Biological duplicates were performed for all growth experiments with at least a 3-day interval between replicate experiments. Cultures were grown in 700 ml of medium. Temperature and pH were maintained at 37 °C and 7.0 throughout the duration of the experiment by PI (proportional - integral) negative feedback controllers. Oxygen availability for aerobic respiration was maintained at atmospheric oxygen level (~ 21%) by pumping in filtered air and monitored by a dissolved O_2 sensor. When O_2 demand due to cellular respiration exceeded oxygen supply in air, 100% oxygen was released into the bioreactor from a compressed O_2 tank by an automated valve on a negative feedback loop. Aeration and mixing was facilitated by a PI feedback controlled impeller running at 250 rpm. Density of the culture was continually monitored by a spectrophotometer (Spectronic 20D+) set to

measure absorbance at 600 nm, using an inline sampling system driven by a peristaltic pump. Cell samples were collected at time points t1 (OD_{600} = 0.2) and t2 (OD_{600} = 0.8) under sterile condition using the sampling port, taking advantage of the positive pressure inside the bioreactor. Equal cell numbers were collected from each time point and spun down in a refrigerated centrifuge at 5000 g. Cell pellets were frozen at -20 °C until plasmid extraction.

Illumina sequencing library preparation

Plasmids were extracted (Sigma Genelute) from frozen cell pellets corresponding to each time point for every growth condition. Oligonucleotides NPPS_CR_FU001-014 were individually mixed with NPPS_CR_FD006_v2 equimolarly to 10 µM, heated to 95 °C for 2 minutes and then allowed to cool to room temperature to form double stranded DNA adapters. 100 ng of the purified plasmid was mixed with 2 µl of the adapter and subjected to a restriction digestion (NheI + SpeI) - ligation (T4 DNA ligase) reaction. The reaction mixture was cycled between 37 °C for 5 minutes and 21 °C for 5 minutes for 15 cycles. After the digestion - ligation reaction, the enzymes in the mixture were heat inactivated and the reaction product was cleaned using a PCR clean-up kit (Zymo DNA Clean & Concentrator-5). 20 ng of the adapter-ligated linearized plasmid was used as PCR template to amplify and enrich the gRNA + random barcode region (NPP252 / NPPS_CR_RX001-RX003). After the enrichment PCR, the libraries were cleaned (Zymo DNA Clean & Concentrator-5) and subjected to a second round of PCR (NPP252/NPP254) to add the Illumina TruSeq flowcell primer sequences to the ends of the libraries. All oligonucleotides sequences are provided in Table 1. The PCR products were once again cleaned and individually quantified for mass and library size, after which the libraries were pooled together equimolarly for sequencing. The pooled library was sequenced on an Illumina HiSeq 2500 high output run set to 125 bp read length and paired-end mode at the University of Minnesota Genomics Center.

Data Analysis

Sequencing data obtained as fastq files was demultiplexed using sample-specific barcodes and sequencing adapters were trimmed (bbduk module from JGI BBTools). Reads containing CRISPR gRNA sequences were matched against a reference file of designed gRNA using blastn module from BLAST+ suite. For each growth condition and time point, the read counts for each valid gRNA sequence was calculated. Data was imported into Matlab for all further analyses. Independent colonies containing the same gRNA were distinguished by the random barcode sequence and treated as replicates. Read counts were normalized per library before fitness values were calculated.

$$Fitness\ w = \frac{\ln\left(d\frac{F_{t2}}{F_{t1}}\right)}{\ln\left(d\frac{W_{t2}}{W_{t1}}\right)}$$

where F_{t1} and F_{t2} are the frequencies of a given genotype at the two timepoints, W_{t1} and W_{t2} are the frequencies of the control gRNA, and d is the expansion factor (van Opijnen, Bodi, and Camilli 2009). Fitness for a genotype was calculated relative to the strain carrying the control gRNA that did not have any targets on the chromosome. Lineage specific fitness was calculated for individual barcodes of a gRNA from the t1 and t2 read counts for each lineage. Bulk fitness for a gRNA genotype (i.e. fitness of a genotype without considering lineages) was calculated as above, with $F = \sum f_i$ where f_i is the read frequency of each lineage for a given gRNA. For a gRNA, standard deviation from mean fitness of the lineages was calculated from the individual lineage fitnesses. The error associated with the measurement of bulk fitness (i.e. the standard deviation) was calculated analytically using the error propagation equation

$$std_w = \sqrt{\left(\frac{\partial w}{\partial F_{t1}}\sigma_{F_{t1}}\right)^2 + \left(\frac{\partial w}{\partial F_{t2}}\sigma_{F_{t2}}\right)^2}$$

where the F_{t1} and F_{t2} are the partial derivatives of the fitness equation.

$$\frac{\partial w}{\partial F_{t1}} = {-1}\Big/{F_{t1}\ln\left(d\frac{W_{t2}}{W_{t1}}\right)} \qquad \frac{\partial w}{\partial F_{t2}} = {1}\Big/{F_{t2}\ln\left(d\frac{W_{t2}}{W_{t1}}\right)}$$

Outliers in barcode lineages were identified using ‘isoutlier’ function in Matlab. A lineage is considered outlier if its fitness is more than 3 scaled median absolute deviations away from the median fitness of lineages of a gRNA. *E. coli* gene essentiality data was downloaded from Ecocyc database (Keseler et al. 2017).

Results

Representation in cloned gRNA library varies with dCas9 induction

Guide RNA plasmid libraries were created in *E. coli* under two conditions - either in the presence of aTc which induced dCas9 expression (+aTc library), or without the inducer (-aTc library) (Figure 1B). Presence of the inducer from the moment of plating the transformed cells causes continuous induction of dCas9, leading to constitutive repression (CR) of target genes in the +aTc library. The -aTc library was grown in bioreactors under different media conditions and gene repression was enabled by addition of the inducer at a predetermined cell density (Timed Repression: TR; Timed Repression - Rich media: TR-Rich; No Repression: NR). The CR growth condition had the lowest number of unique gRNA recovered after growth experiments because gRNA-targeted repression is active even during initial colony growth (Figure 2). One would expect that the NR growth condition has the best representation of recovered gRNA because dCas9 was not induced at all and this holds true. TR gRNAs are better represented compared to CR because repression is functional for only a few generations and this might not sufficient to deplete cells targeted for essential genes to below detectable levels. TR-Rich is the second best represented condition because many essential genes in TR will be non-essential under the rich media condition. In conditions in which dCas9 was expressed (CR, TR, TR-Rich), there was high concordance in the number of targeted genes present in both biological replicates. The number of genes uniquely present in one replicate of a growth condition but not in the other was less than 1% of the genes present in common between the replicates. Overall, 3410 genes were represented across all conditions and replicates.

gRNAs can be tagged with barcodes post oligo library synthesis to track individual lineages (colonies) of a gRNA

To track lineages arising from independent colonies of the same gRNA genotype, it was necessary to tag the gRNA coding sequences with barcodes after oligolibrary synthesis. This tagging of gRNA associated lineage barcodes (GLB) was done during the plasmid cloning step (Figure 1A). The median number of barcodes per gRNA and the distribution of barcode numbers across the gRNAs were very similar across the conditions tested (Figure 3), except for the more extreme outliers observed in CR. This could be due to artifacts in PCR amplification or barcode

ligation (note that the +aTc and -aTc libraries were created separately). For most gRNAs, the number of functional lineages as recovered by NGS was consistent between replicates for a given condition (Figure 4). This indicated that there wasn't a widespread dropout or fluctuation in the number of lineages during the growth experiments, attesting to the biological reproducibility of target gRNA function.

GLBs enable lineage-specific fitness analysis and act as internal replicates to uncover inconsistencies in growth experiments

Fitness for a gRNA genotype was calculated in two ways. First, for each barcoded lineage of a gRNA, fitness was calculated based on the frequencies of the lineage at t1 & t2. The individual fitness value of each lineage was averaged to find the mean barcode fitness (MBF) for each gRNA. Next, the barcodes were ignored and the read frequencies at t1 & t2 for each gRNA (i.e. sum of lineage frequencies for a gRNA) were used to calculate the bulk fitness (BF) of the gRNA. We compared MBF and BF for a replicate and also between replicates of a growth condition to check for correlations (Figure 5) We found that MBF and BF were almost identical for a given replicate (Pearson's $\rho > 0.99$). However, between replicates, there was a positive but weaker correlation in the fitness values. Both MBF and BF had similar correlation values in pairwise comparisons between replicates. The weaker correlation could be due to stochasticity in the various components of the experimental system, since we performed the replicate experiments to be as identical as possible.

The error associated with the fitness measurement or the standard deviation (SD) on fitness was explored to examine how consistently the independent lineages of a gRNA reported on the fitness. There was a very weak negative correlation between barcode lineage SD (LSD) and SD estimated for the bulk fitness (BSD) for a given replicate. Variation in fitness values between lineages of a genotype can be due to biological stochasticity and measurement noise. As it is determined from observation, LSD captures this variation and is a better validation of fitness compared to BSD which can only be estimated from the bulk frequencies of genotypes. Indeed, there was a weak positive correlation in LSD between replicates.

For a given gRNA, fitness of some lineages can be outliers due to several reasons: selection pressure relieving the repression of dCas9, a stochastic change in expression strength of another gene that unexpectedly alleviates fitness, noise and biased in the molecular methods for NGS

etc. gRNA that had at least 10 lineages were considered for lineage outlier analysis. Approximately 30% of the genes that were analyzed per replicate had at least one outlier that could be removed (Figure 6A). In some cases, the removal of an outlier lineage produced a drastic change in the fitness (MBF) calculated for the gRNA (Figure 6B). Corrected MBF of a gRNA after removal of outlier lineages was predominantly within 0.05 units of initial MBF (Figure 7)

Timed Repression has a weaker effect on fitness compared to Constitutive Repression

In timed repression, gRNA that target essential genes can be included in the library without affecting the phenotype. After expansion of these genotypes without any gene repression, dCas9 can be induced to repress the target gene and the subsequent drop in fitness observed. Such an assay is useful to confirm if a gene is indeed essential and affects the organism's fitness. In transposon screens, essentiality is inferred rather than confirmed because inactivating an essential gene makes the strain unviable, resulting in no colonies.

Gene fitnesses (mean fitness of lineages) were compared between TR and CR conditions to check if they were comparable between the two repression schemes (Figure 8). A significant number of essential genes were detected in the TR condition but not in the CR condition, indicating that constitutive repression of essential genes leads to unviable cells. Further, the essential genes detected in TR but not in CR spanned the full range in fitness. The period of repression could be insufficient to fully negate the essential gene activity i.e. enough time might not be available to completely repress the transcription of the targeted gene and deplete the gene products by RNA and protein turnover to have a strong negative effect on the phenotype. Non-essential genes that were in the low fitness range (arbitrarily defined as < 0.5) had a higher fitness value in TR compared to CR. Overall, timed repression had a weaker effect on fitness but can be a useful strategy to directly confirm gene essentiality, particularly if more samples are collected during a longer repression period. In such an experiment, the depletion of gRNA targeting essential genes can be experimentally tracked.

Timed Repression enables assessing the same library under multiple media conditions

Timed repression enables assaying the same starter library under multiple conditions, allowing testing of conditional gene essentiality (where a gene essential in one condition is non-essential in another). Studying conditional essentiality is labor-intensive in cases of constitutively repressed CRISPRi, transposon mutagenesis, and targeted gene deletion, as the strain libraries need to be created anew for each condition that is being tested. Timed repression combined with phenotype microarrays (Bochner, Gadzinski, and Panomitros 2001) can trivially determine conditional gene essentiality across a wide range of growth substrates using the same starting library of genotypes.

Gene representation was compared between TR and TR-Rich conditions. Genes that play a role in biosynthesis were depleted in TR minimal media but present in TR-Rich (Table 2). Nearly 1/3rd of the genes that were present only in TR-Rich had been identified as essential, and another 1/3rd had been identified as ambiguous (deletions that could not be confirmed for removal of gene (Yamamoto et al. 2009)).

Discussion

CRISPR-based interference screens are a new paradigm in bacterial transcriptome control. There is no prior technology available in bacteria that works at a similar scale, and thus there is no established baseline for comparison of CRISPRi performance. Using gRNA associated linear barcodes, we were able to quantify the variability in CRISPRi effectiveness within isogenic strains. We could also uncover lineages whose fitness values were outliers compared to the rest of the lineages for a gRNA and remove these outliers for more accurate fitness measurements.

Because CRISPRi is a repression system that doesn't permanently disrupt a target gene, we expect both gRNA design and properties of the target gene (expression strength, genomic location) to influence the efficiency of repression. This is analogous to Tn-Seq, where insertions in different locations within a gene can produce different fitness phenotypes (Yang et al. 2014; Goodall et al. 2018). We designed only one gRNA per gene due to limitations of the synthesis technology we used to manufacture our oligo library. To improve the accuracy of fitness measurements and minimize variation in repression efficiency, one can design multiple gRNA that tile the entirety of a gene sequence (Tianmin Wang et al. 2017).

CRISPRi is a very flexible assay for exploring gene function. The number of targets probed by the assay can be easily varied from a single gene to the genome-scale, all while being amenable to phenotyping by NGS. Unlike transposon mutagenesis, CRISPRi can be focused on specific sets of gene targets that are of interest. It is trivial to target only the genes of a biochemical subsystem like lipid metabolism or secondary metabolite biosynthesis. Desired subsets of gRNA can be amplified from a pooled genome-wide gRNA oligo library by using the principles of dial-out PCR (Schwartz, Lee, and Shendure 2012), enabling exquisite customization and rapid creation of CRISPRi libraries. Of great interest is the targeting of essential genes during cell growth. Insertional or targeted mutations in essential genes lead to cell death, resulting in these genes not being represented in functional genomics studies. However, essential genes are of significant interest precisely because of their essentiality. Past efforts to resolve this catch-22 situation relied on introducing a plasmid-borne copy of the essential gene into the organism, and then deleting the chromosomal copy by recombination. Either plasmid replication or essential gene induction was made conditional. This complicated process is highly simplified by CRISPRi, where 1) essential genes are targeted in the same way as any non-essential gene, 2) temporal control is available over gene repression 3) degree of repression can be tuned by controlling the expression of dCas9. CRISPRi is singularly useful in probing gene essentiality in less studied organisms (Peters et al. 2016). Because CRISPRi can target sequences with precision, it is highly suited for repressing genes encoding small ncRNAs and mini-protein, targets which are small enough to escape transposon insertions. In this study, we explored the use of CRISPRi libraries for single gene repression. CRISPRi can be expanded to study genetic interactions by cloning two or more gRNA as an array. This approach works best for probing small gene sets for genetic interactions as performing a genome-scale GI screen can be technically challenging (as there is a requirement to create (N*(N-1))/2 unique colonies by direct transformation in an organism with N genes).

In addition to the widely used dCas9 from *Streptococcus pyogenes*, orthogonal versions are now available that exhibit different PAM requirements which are useful in organisms with skewed genome base composition (Esvelt et al. 2013). Other types of sequence-guided genome editing tools are also gaining traction. Cpf1 is an RNA-guided DNA endonuclease that requires only the gRNA for its function, foregoing the additional noncoding RNA required by Cas9 (Zetsche et al. 2015; S. K. Kim et al. 2017). Such compactness would be an advantage in CRISPRi experiments where strong selection is present against the activity of the CRISPRi components. A more notable discovery is that of Cas13a (C2C2), which is an RNA-guided RNA nuclease that can be used for direct RNA knockdowns (Abudayyeh et al. 2017). These advances make CRISPRi a

powerful tool with wide application. CRISPRi is more targeted than transposons and more cost-effective than targeted deletion libraries, thus meriting serious consideration for functional genomics screens. We have shown that barcodes for lineage analysis improve quality of CRISPRi measurement, and inducible expression of dCas9 provides temporal flexibility in expression control. We expect that the innovations explored here will enhance the value of CRISPRi as a functional genomics tool in microbiological studies.

Figures

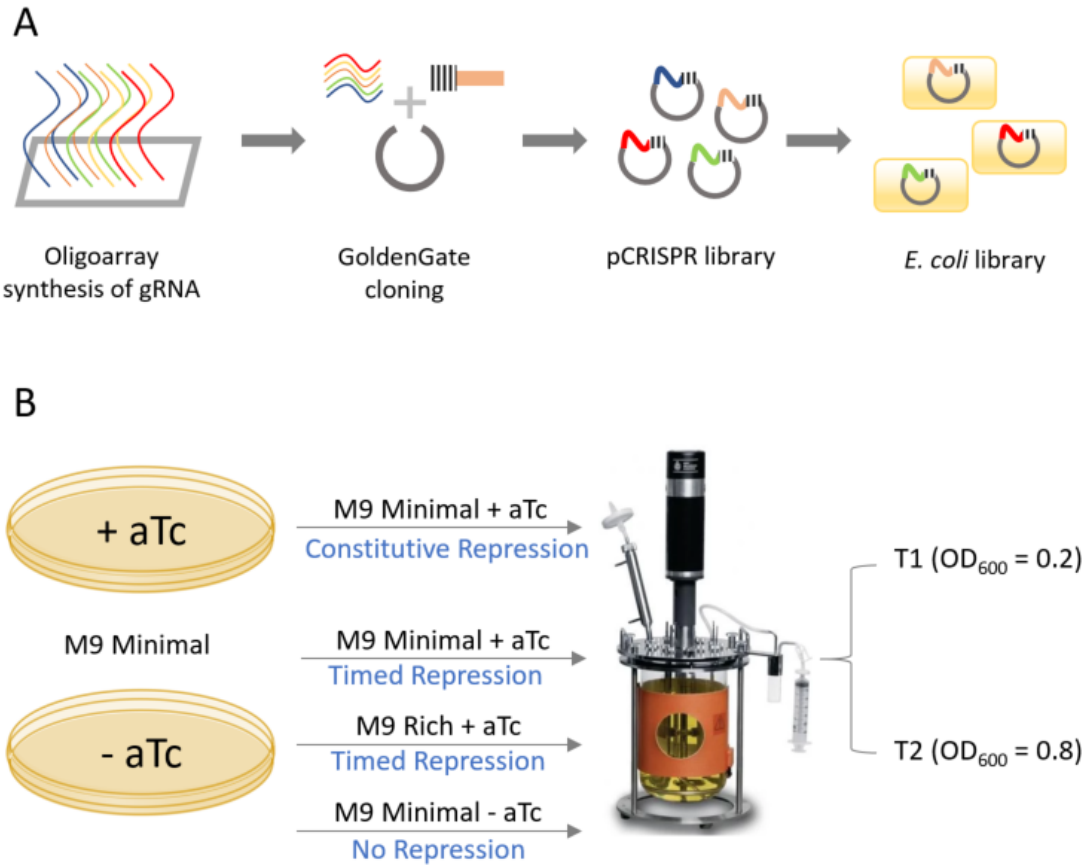


Figure 1: CRISPRi library creation and growth experiments A) gRNA for genome-wide targeting were synthesized as an oligolibrary and cloned *en masse* in combination with DNA barcodes into a plasmid vector. The ligated plasmid library was transformed into *E. coli* DH5α: (pdcas9-bacteria) by electroporation. B) Transformed bacteria were plated either in the presence or absence of anhydrotetracycline to obtain libraries that had constitutively repression (+aTc) or had no repression (-aTc) activity respectively. Libraries were then grown in an environment-controlled bioreactor under different media and dCas9 induction conditions, with cell samples collected at two points for NGS based fitness analysis.

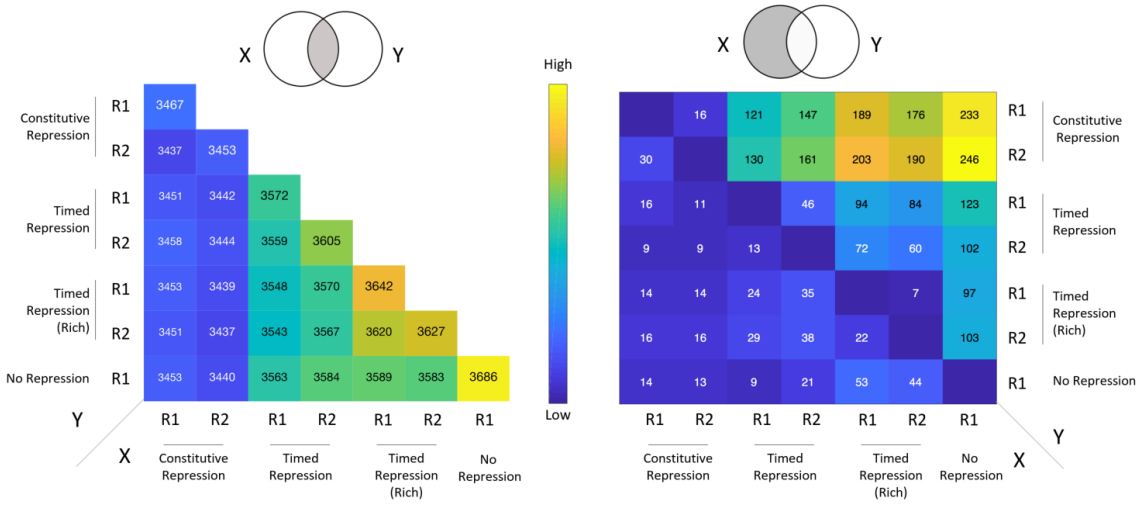


Figure 2: CRISPR gRNA identified across growth conditions To compare the performance of CRISPRi across replicates and experiments, it was necessary that a substantial number of gRNA were identified by NGS across all libraries. R1 and R2 indicate biological replicates. Replicates of constitutive repression had the least number of shared gRNA sequences. They also were missing the most gRNA compared to other conditions. As expected, more gRNA were identified in the timed repression condition compared to constitutive repression, and a similar trend was observed when rich media was compared to minimal media. The low number of unique gRNA present in one replicate but not in the other (less than 1%) indicated that the replicates would be useful for further comparative analysis.

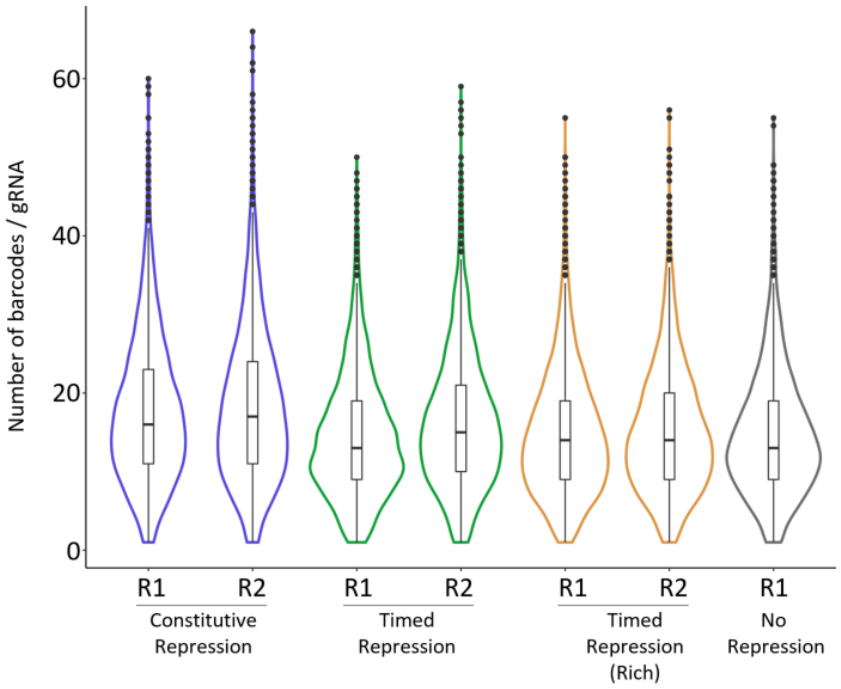


Figure 3: Distribution of gRNA associated lineage barcodes across experiments The median number of barcodes recovered per gRNA is fairly consistent across experiments (box plot), even if the constitutive repression library has more extreme outliers. Also, the shape of the probability density distribution shows that there are very few gRNA with abnormally high numbers of barcodes associated with them. The variation in barcodes per gRNA could be due to variation in gRNA abundance within the PCR amplified oligolibrary used for cloning.

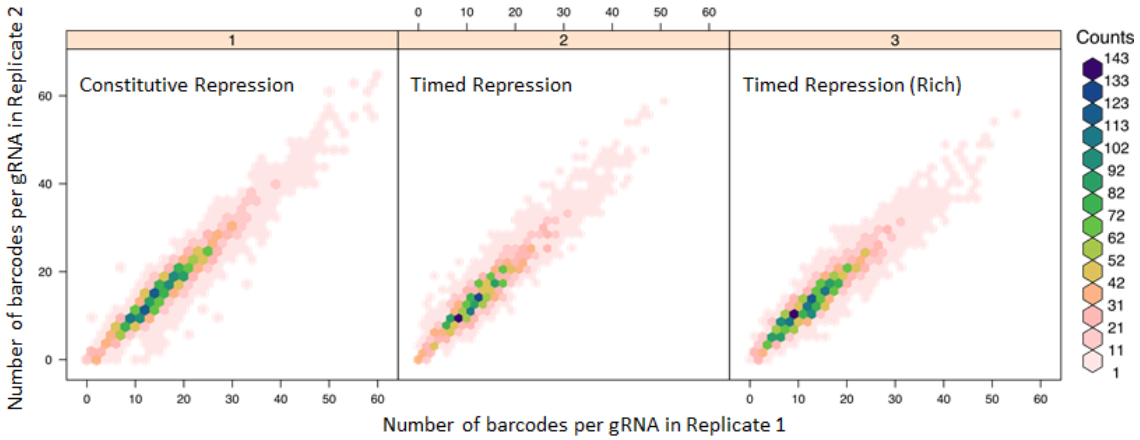


Figure 4: GLB concordance across replicates For each gRNA that was detected in both replicates of a condition, the number of barcodes recovered in replicate 1 was plotted against the number of barcodes recovered in replicate 2. Each hexagonal bin in the plot is an intersection of points of the two axes (scatter plot) and is shaded by the count of gRNA that have the same intersection point (histogram). It is apparent from the figure that for most gRNA, a similar number of barcodes are recovered between replicates. Any deviation could be due to sampling issues during growth experiments or NGS data (reads containing a particular barcode) that was discarded due to sequence errors.

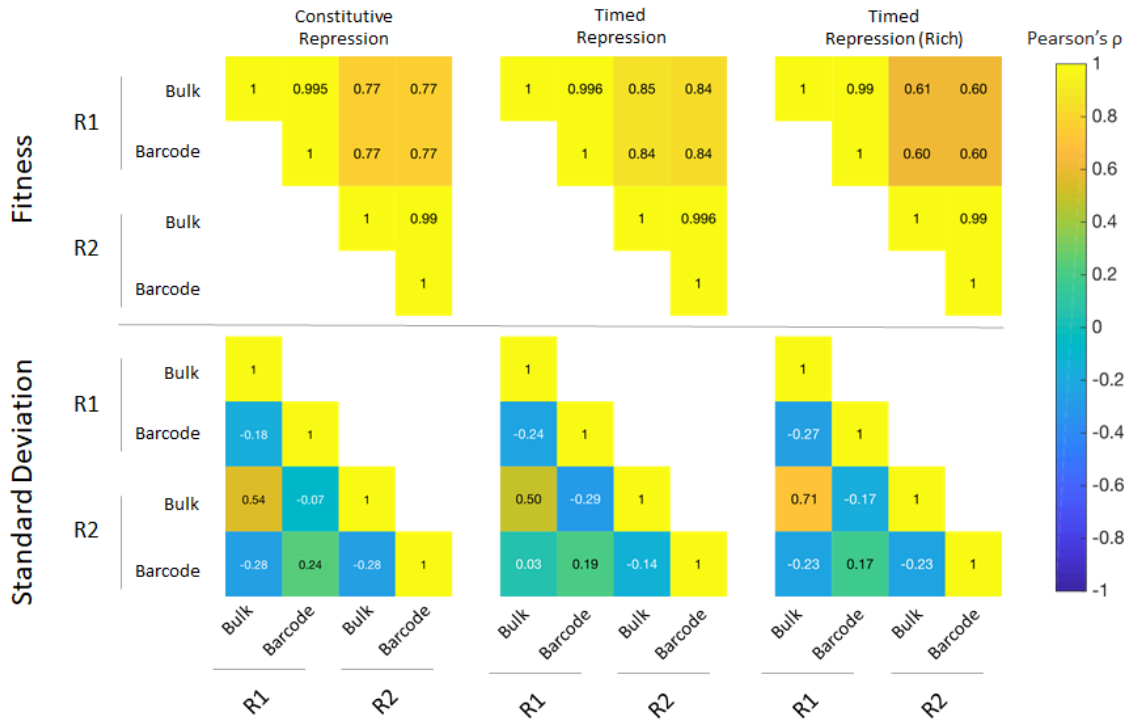
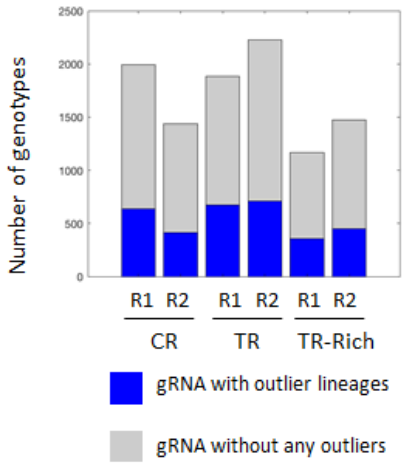


Figure 5: Fitness and standard deviation correlation between biological replicates Within a replicate, bulk fitness and mean barcode fitness for a gRNA are very highly correlated, indicating that gRNA fitness derived as the average of barcode fitness is almost identical to the conventionally calculated gRNA fitness. Between-replicate correlations of either bulk or barcode derived fitness are positive but less strong which could be due to variations introduced by stochasticity on the days the experiments were performed, even if the experiments were replicated accurately as humanly possible. There is a slight negative correlation between bulk SD and barcode SD within a replicate. While the bulk SD is analytically estimated from gRNA frequencies at each sampling timepoint, the barcode SD is calculated from the observed fitnesses of barcodes for every gRNA and is therefore expected to more accurately reflect the error associated with fitness measurements. This is supported by the weak positive correlation of barcode SD between replicates.

A



B

minD gRNA in CR – Replicate 1

Lineage	Read frequency t1	Read frequency t2	Fitness
1	618	176	-0.15
2	227	295	0.90
3	213	317	1.00
4	200	269	0.93
5	181	241	0.92
6	160	251	1.03
7	139	192	0.95
8	125	192	1.02
9	117	157	0.93
10	105	160	1.01
11	101	165	1.06
12	74	125	1.09

Bulk Fitness = 0.80

Mean Barcode Fitness = 0.89

Mean Barcode Fitness after removing outlier = 0.98

Figure 6: Analysis for outlier lineages Genotypes that had at least 10 lineages were analyzed for outliers in calculated fitness. A) Approximately 30% of the genotypes analyzed contained outliers that could be removed. R1 & R2 are replicates 1 & 2. B) Example analysis: The minD gRNA in Constitutive Repression condition - Replicate 1 had 12 lineages. However, one lineage was a clear outlier with a negative fitness value, resulting from t2 < t1. The higher reads in t1 or the lower reads in t2 could have been due to PCR or NGS errors. Removing the outlier substantially changes the fitness of the genotype. This analysis would not have been possible with bulk fitness measurement, thus reporting an incorrect phenotype for this genotype.

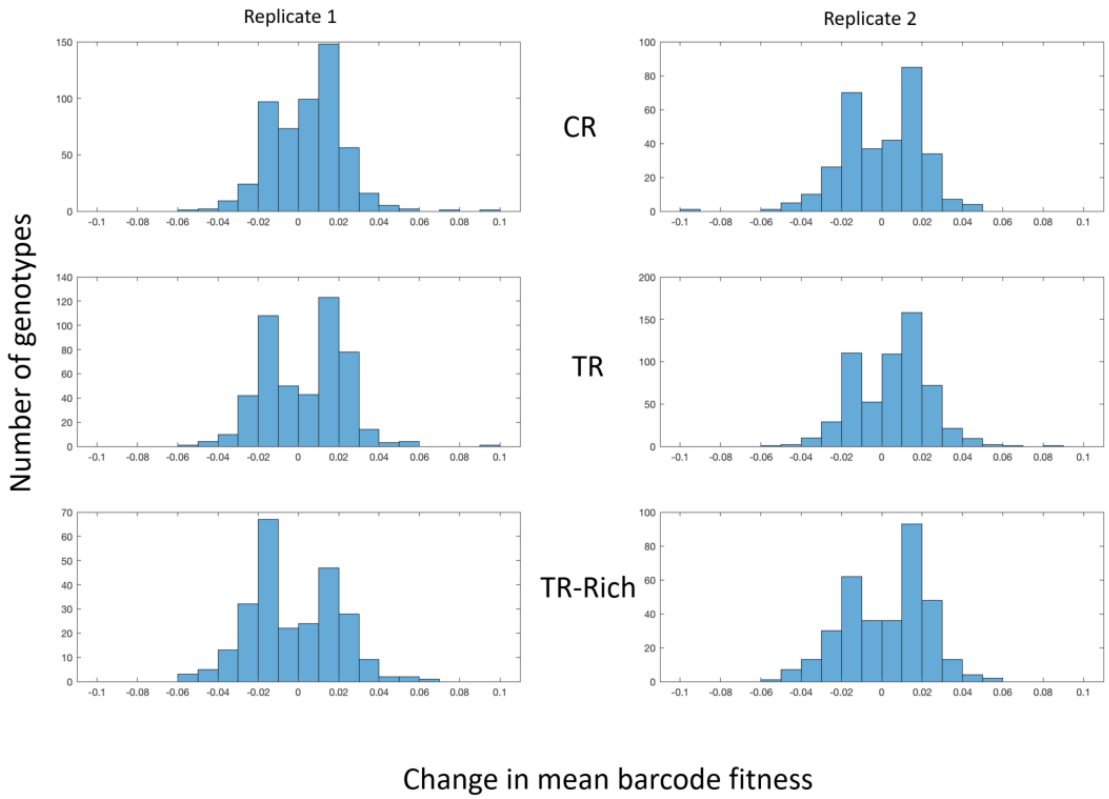


Figure 7: Histogram of change in fitness after removal of outlier lineages The histograms show how the corrected mean barcode fitness (cMBF) changed with respect to MBF after removing lineages with anomalous fitness values. Most of the changes were within 0.05 units of the MBF value.

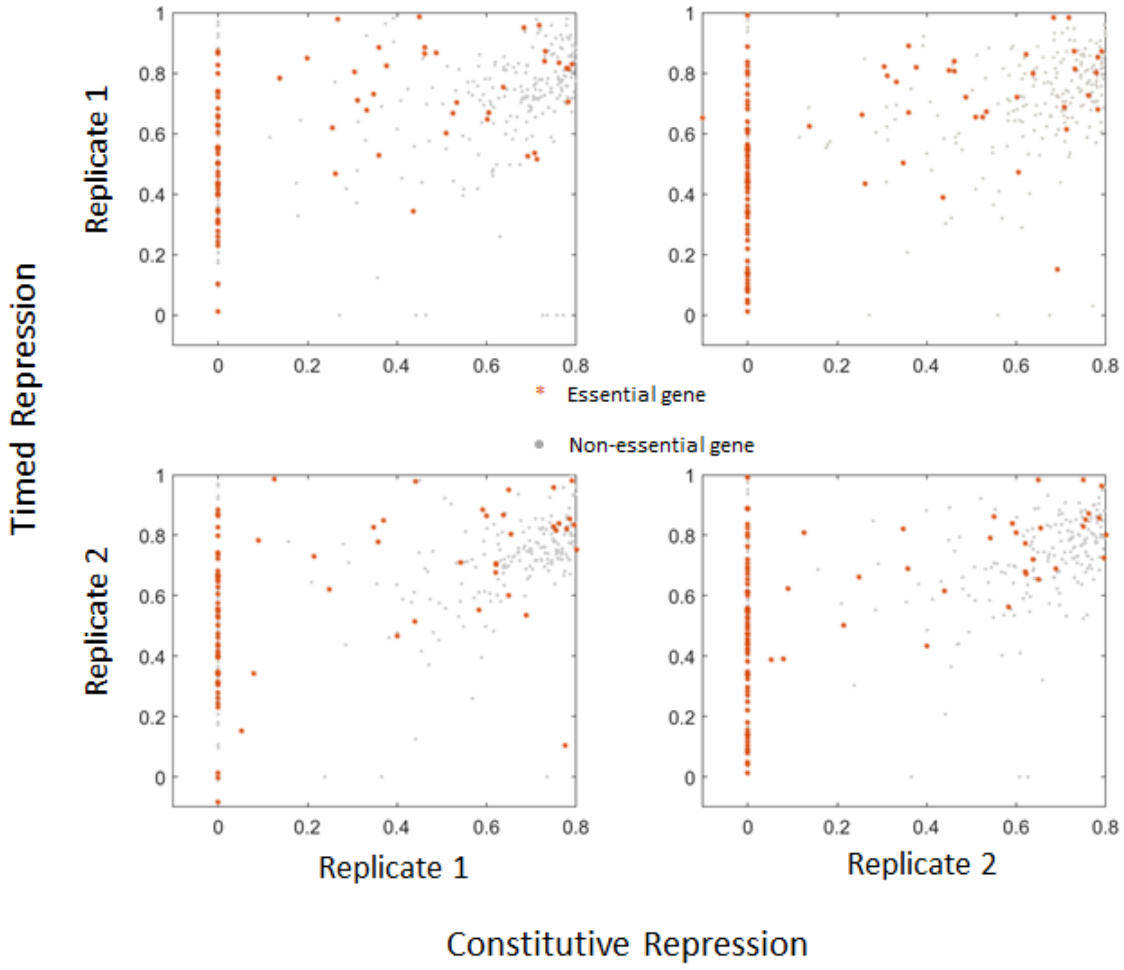


Figure 8: Comparison of Timed Repression to Constitutive Repression For each gRNA genotype detected in TR or CR condition, fitness (mean fitness of barcoded lineages) was compared between the two conditions. If a gRNA was not detected in one of the conditions, then the fitness value for that genotype was set to 0 in that condition. The axes of the scatter plots are fitness values. A substantial number of gRNA for essential genes are detected in TR whereas these are not found in the CR (see y-intercept in the plots). Further, for the essential genes detected in TR but not in CR, the fitness values span a wide range. This indicates that the period of timed repression was not sufficient to completely deplete the essential gene's activity. The axes limits were chosen to display the most informative part of the figures.

Tables

Oligonucleotide Name	Sequence
NPP260	CTTGGTCAGACGAGTGCATGG
NPP261	GACCGGCAATCTCTTCCTGG
NPP342	NNNGGTCTCN GCTGTTTTGAATGGTCCCAAAAC NNNN NNNN NNNN TTCAGCACACTGAGACTTGTTGAG
NPP397	NNNNGGTCTCN TCTG GGATACACCAAGGAAAGTCTACACGAAC C
NPP420	GTCTAAAGTATGCGTCGCGGCATG GTCTCN CAG ATTCAGGACGAGCCTCAGACTC
NPP421	CGGGATATGGGGCTCCTTTAGCGAC GCTGAATTGTGGTGGACGAATTCTC
NPPS_CR_FU001 - NPPS_CR_FU014	CCCTACACGACGCTCTTCCGATCT X YYYY GCCACCTTAACACGCGATGAG where X YYYY is NN AGCC NNN GCCT NNNN CGAG NNNNN TTAC NNN TAGA NNN GCAA NN GAAT N ACAT NN GTCG NNN ATGT NNNN ATTG N CGCA NN TGCG NNN CCTA
NPPS_CR_FD006_v2	/5Phos/CTAGCTCATCGCGTGTTAAGGTGGC
NPPS_CR_RX001	GAGTTCAGACGTGTGCTCTTCCGATCT NN CTCAACAAGTCTCAGTGTGCTG
NPPS_CR_RX002	GAGTTCAGACGTGTGCTCTTCCGATCT NNN CTCAACAAGTCTCAGTGTGCTG
NPPS_CR_RX003	GAGTTCAGACGTGTGCTCTTCCGATCT NNNN CTCAACAAGTCTCAGTGTGCTG

NPP252	AATGATACGGCGACCACCGAGATCTACACTCTTTC CCTACACGACGCTCTTCCGATCT
NPP254	CAAGCAGAAGACGGCATACGAGAT CCGAGAT GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Table 1: List of oligonucleotides used in this study

Non-essential	Essential	Ambiguous
ampD	murC	lpd
mhpR	dapD	bioA
yajC	tsf	bioF
cstA	lpxH	bioD
pgm	lnt	aroD
sucA	ftsK	hisC
sucB	aroA	hisB
aspC	msbA	hisF
putA	lpxK	yfaD
rssB	asnS	aroC
hns	murJ	nadB
tyrR	fabG	cysN
nudB	thrS	cysH
preA	gapA	cysJ
ccmD	dapE	lysA
cysK	rplS	serA
cysW	pgk	metC
cysU	rpoD	argG
cysP	nusA	aroE
purN	folP	panM
srlE	rpmA	pitA
rppH	lptB	ilvE
rdgB	degS	ilvD
folB	tsaC	metE
ptsN	fmt	rfaH
aroK	rpsH	rhaB
waaF	yidC	pfkA

ilvN	glmU	serB
ilvB	murI	
ivbL	rpoB	
yidR	rpoC	
trxA		
thiC		
cysQ		

Table 2: Genes identified in TR-Rich but not in TR (minimal media).

Appendix

Sequence of synthesized DNA fragment (gBlock from IDT) encoding tracrRNA and promoter for gRNA

```
GTCGCTAAAGGAGCCCCATATCCCGTTACTATAAGCCTATTGAGTATTTCTTATCCATTTTTGCCTCCTAA
AATAAAAAGTTTAAATTAAATCCATAATGAGTTTGATGATTTCAATAATAGTTTTAATGACCTCCGAAATT
AGTTTAATATGCTTTAATTTTTCTTTTTCAAAATATCTCTTCAAAAAATATTACCCAATACTTAATAATAA
ATAGATTATAACACAAAATTCTTTTAAAAAGTAGTTTATTTTGTTATCATTCTATAGTATTAAGTATTGTT
TTATGGCTGATAAATTTCTTTGAATTTCTCCTTGATTATTTGTTATAAAAGTTATAAAATAATCTTGTTGG
AACCATTCAAAACAGCATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGT
CGGTGCTTTTTTTGATACTTCTATTCTACTCTGAGTATATTTTAGATGAAGATTATTTCTTAATAACTAAA
AATATGGTATAATACTCTTAATAAATGCAGTAATACAGGGGCTTTTCAAGACTGAAGTCTAGCTGAGACAA
ATAGTGCGATTACGAAATTTTTTAGACAAAAATAGTCTACGAGGTTTTAGAGCTATGCTGTTTTGAATGGT
CCCAAAACTGAGACCAGTCTCGGAAGCGTCTAAAGTATGCGTCGCGGCATG
```

Bibliography

Abudayyeh, Omar O., Jonathan S. Gootenberg, Patrick Essletzbichler, Shuo Han, Julia Joung, Joseph J. Belanto, Vanessa Verdine, et al. 2017. “RNA Targeting with CRISPR-Cas13.” *Nature* 550 (7675): 280–84.

Aidelberg, Guy, Benjamin D. Towbin, Daphna Rothschild, Erez Dekel, Anat Bren, and Uri Alon. 2014. “Hierarchy of Non-Glucose Sugars in Escherichia Coli.” *BMC Systems Biology* 8 (December): 133.

Akerley, B. J., E. J. Rubin, A. Camilli, D. J. Lampe, H. M. Robertson, and J. J. Mekalanos. 1998. “Systematic Identification of Essential Genes by in Vitro Mariner Mutagenesis.” *Proceedings of the National Academy of Sciences of the United States of America* 95 (15): 8927–32.

Allen, Doug K., Joshua Goldford, James K. Gierse, Dominic Mandy, Christine Diepenbrock, and Igor G. L. Libourel. 2014. “Quantification of Peptide M/z Distributions from 13C-Labeled Cultures with High-Resolution Mass Spectrometry.” *Analytical Chemistry* 86 (3): 1894–1901.

Anders, Carolin, Ole Niewoehner, Alessia Duerst, and Martin Jinek. 2014. “Structural Basis of PAM-Dependent Target DNA Recognition by the Cas9 Endonuclease.” *Nature* 513 (7519): 569–73.

Antoniewicz, Maciek R., Joanne K. Kelleher, and Gregory Stephanopoulos. 2007. “Elementary Metabolite Units (EMU): A Novel Framework for Modeling Isotopic Distributions.” *Metabolic Engineering* 9 (1): 68–86.

Auffray, Charles, Sandrine Imbeaud, Magali Roux-Rouquié, and Leroy Hood. 2003. “From Functional Genomics to Systems Biology: Concepts and Practices.” *Comptes Rendus Biologies* 326 (10-11): 879–92.

Baba, Tomoya, Takeshi Ara, Miki Hasegawa, Yuki Takai, Yoshiko Okumura, Miki Baba, Kirill A. Datsenko, Masaru Tomita, Barry L. Wanner, and Hirotada Mori. 2006. “Construction of Escherichia Coli K-12 in-Frame, Single-Gene Knockout Mutants: The Keio Collection.” *Molecular Systems Biology* 2 (February): 2006.0008.

Babu, Mohan, J. Javier Díaz-Mejía, James Vlasblom, Alla Gagarinova, Sadhna Phanse, Chris Graham, Fouad Yousif, et al. 2011. “Genetic Interaction Maps in Escherichia Coli Reveal Functional Crosstalk among Cell Envelope Biogenesis Pathways.” *PLoS Genetics* 7 (11): e1002377.

Baird, G. S., D. A. Zacharias, and R. Y. Tsien. 2000. “Biochemistry, Mutagenesis, and Oligomerization of DsRed, a Red Fluorescent Protein from Coral.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (22): 11984–89.

Bandaru, Pradeep, Neel H. Shah, Moitrayee Bhattacharyya, John P. Barton, Yasushi Kondo, Joshua C. Cofsky, Christine L. Gee, et al. 2017. “Deconstruction of the Ras Switching Cycle through Saturation Mutagenesis.” *eLife* 6 (July). https://doi.org/10.7554/eLife.27810.

Barnett, Derek W., Erik K. Garrison, Aaron R. Quinlan, Michael P. Strömberg, and Gabor T. Marth. 2011. “BamTools: A C++ API and Toolkit for Analyzing and Managing BAM Files.” *Bioinformatics* 27 (12): 1691–92.

Barrangou, Rodolphe, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A. Romero, and Philippe Horvath. 2007. “CRISPR Provides Acquired Resistance against Viruses in Prokaryotes.” *Science* 315 (5819): 1709–12.

Baryshnikova, Anastasia, Michael Costanzo, Chad L. Myers, Brenda Andrews, and Charles Boone. 2013. “Genetic Interaction Networks: Toward an Understanding of Heritability.” *Annual Review of Genomics and Human Genetics* 14 (June): 111–33.

Bernards, René, Thijn R. Brummelkamp, and Roderick L. Beijersbergen. 2006. “shRNA Libraries and Their Use in Cancer Genetics.” *Nature Methods* 3 (9): 701–6.

Bevis, Brooke J., and Benjamin S. Glick. 2002. “Rapidly Maturing Variants of the Discosoma Red Fluorescent Protein (DsRed).” *Nature Biotechnology* 20 (1): 83–87.

Bikard, David, Wenyan Jiang, Poulami Samai, Ann Hochschild, Feng Zhang, and Luciano A. Marraffini. 2013. “Programmable Repression and Activation of Bacterial Gene Expression

Using an Engineered CRISPR-Cas System." *Nucleic Acids Research* 41 (15): 7429–37.
Bochner, B. R., P. Gadzinski, and E. Panomitros. 2001. "Phenotype Microarrays for High-Throughput Phenotypic Testing and Assay of Gene Function." *Genome Research* 11 (7): 1246–55.
Bowater, Richard, and Aidan J. Doherty. 2006. "Making Ends Meet: Repairing Breaks in Bacterial DNA by Non-Homologous End-Joining." *PLoS Genetics* 2 (2): e8.
Bowden, Steven D., Nagendra P. Palani, and Igor G. L. Libourel. 2017. "Stringent Control of FLP Recombinase in Escherichia Coli." *Journal of Microbiological Methods* 133 (February): 52–54.
Brutinel, Evan D., and Jeffrey A. Gralnick. 2012. "Anomalies of the Anaerobic Tricarboxylic Acid Cycle in Shewanella Oneidensis Revealed by Tn-Seq." *Molecular Microbiology* 86 (2): 273–83.
Buchholz, F., L. Ringrose, P. O. Angrand, F. Rossi, and A. F. Stewart. 1996. "Different Thermostabilities of FLP and Cre Recombinases: Implications for Applied Site-Specific Recombination." *Nucleic Acids Research* 24 (21): 4256–62.
Bulat, Sergey A. 2016. "Microbiology of the Subglacial Lake Vostok: First Results of Borehole-Frozen Lake Water Analysis and Prospects for Searching for Lake Inhabitants." *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 374 (2059). https://doi.org/10.1098/rsta.2014.0292.
Burdge, David A., and Igor G. L. Libourel. 2014. "Open Source Software to Control Bioflo Bioreactors." *PloS One* 9 (3): e92108.
Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (December): 421.
Campbell, Robert E., Oded Tour, Amy E. Palmer, Paul A. Steinbach, Geoffrey S. Baird, David A. Zacharias, and Roger Y. Tsien. 2002. "A Monomeric Red Fluorescent Protein." *Proceedings of the National Academy of Sciences of the United States of America* 99 (12): 7877–82.
Check Hayden, Erika. 2014. "Technology: The $1,000 Genome." *Nature News* 507 (7492): 294.
Chen, Xuewen, Ana P. Alonso, Doug K. Allen, Jennifer L. Reed, and Yair Shachar-Hill. 2011. "Synergy between (13)C-Metabolic Flux Analysis and Flux Balance Analysis for Understanding Metabolic Adaptation to Anaerobiosis in E. Coli." *Metabolic Engineering* 13 (1): 38–48.
Chi, Xu, Yingchun Zhang, Zheyong Xue, Laibao Feng, Huaqing Liu, Feng Wang, and Xiaoquan Qi. 2014. "Discovery of Rare Mutations in Extensively Pooled DNA Samples Using Multiple Target Enrichment." *Plant Biotechnology Journal* 12 (6): 709–17.
Choi, Young J., Lyne Morel, Teffanie Le François, Denis Bourque, Lucie Bourget, Denis Groleau, Bernard Massie, and Carlos B. Míguez. 2010. "Novel, Versatile, and Tightly Regulated Expression System for Escherichia Coli Strains." *Applied and Environmental Microbiology* 76 (15): 5058–66.
Cong, Le, F. Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D. Hsu, et al. 2013. "Multiplex Genome Engineering Using CRISPR/Cas Systems." *Science* 339 (6121): 819–23.
Conway, Tyrrell, James P. Creecy, Scott M. Maddox, Joe E. Grissom, Trevor L. Conkle, Tyler M. Shadid, Jun Teramoto, et al. 2014. "Unprecedented High-Resolution View of Bacterial Operon Architecture Revealed by RNA Sequencing." *mBio* 5 (4): e01442–14.
Costanzo, Michael, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D. Spear, Carolyn S. Sevier, Huiming Ding, et al. 2010. "The Genetic Landscape of a Cell." *Science* 327 (5964): 425–31.
Costanzo, Michael, Benjamin VanderSluis, Elizabeth N. Koch, Anastasia Baryshnikova, Carles Pons, Guihong Tan, Wen Wang, et al. 2016. "A Global Genetic Interaction Network Maps a Wiring Diagram of Cellular Function." *Science* 353 (6306). https://doi.org/10.1126/science.aaf1420.
Côté, Jean-Philippe, Shawn French, Sebastian S. Gehrke, Craig R. MacNair, Chand S. Mangat,

Amrita Bharat, and Eric D. Brown. 2016. “The Genome-Wide Interaction Network of Nutrient Stress Genes in Escherichia Coli.” *mBio* 7 (6). https://doi.org/10.1128/mBio.01714-16.
Datsenko, K. A., and B. L. Wanner. 2000. “One-Step Inactivation of Chromosomal Genes in Escherichia Coli K-12 Using PCR Products.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (12): 6640–45.
Dauner, M., and U. Sauer. 2000. “GC-MS Analysis of Amino Acids Rapidly Provides Rich Information for Isotopomer Balancing.” *Biotechnology Progress* 16 (4): 642–49.
DeJesus, Michael A., Subhalaxmi Nambi, Clare M. Smith, Richard E. Baker, Christopher M. Sassetti, and Thomas R. Ioerger. 2017. “Statistical Analysis of Genetic Interactions in Tn-Seq Data.” *Nucleic Acids Research* 45 (11): e93.
Dixit, Atray, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, et al. 2016. “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens.” *Cell* 167 (7): 1853–66.e17.
Engler, Carola, Romy Kandzia, and Sylvestre Marillonnet. 2008. “A One Pot, One Step, Precision Cloning Method with High Throughput Capability.” *PloS One* 3 (11): e3647.
Esvelt, Kevin M., Prashant Mali, Jonathan L. Braff, Mark Moosburner, Stephanie J. Yaung, and George M. Church. 2013. “Orthogonal Cas9 Proteins for RNA-Guided Gene Regulation and Editing.” *Nature Methods* 10 (11): 1116–21.
Farmer, Hannah, Nuala McCabe, Christopher J. Lord, Andrew N. J. Tutt, Damian A. Johnson, Tobias B. Richardson, Manuela Santarosa, et al. 2005. “Targeting the DNA Repair Defect in BRCA Mutant Cells as a Therapeutic Strategy.” *Nature* 434 (7035): 917–21.
Firth, Andrew E., and Wayne M. Patrick. 2005. “Statistics of Protein Library Construction.” *Bioinformatics* 21 (15): 3314–15.
Fowler, Douglas M., Carlos L. Araya, Sarel J. Fleishman, Elizabeth H. Kellogg, Jason J. Stephany, David Baker, and Stanley Fields. 2010. “High-Resolution Mapping of Protein Sequence-Function Relationships.” *Nature Methods* 7 (9): 741–46.
Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. “CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data.” *Bioinformatics* 28 (23): 3150–52.
Ganti, Tibor. 2003. *The Principles of Life*. OUP Oxford.
Garst, Andrew D., Marcelo C. Bassalo, Gur Pines, Sean A. Lynch, Andrea L. Halweg-Edwards, Rongming Liu, Liya Liang, et al. 2017. “Genome-Wide Mapping of Mutations at Single-Nucleotide Resolution for Protein, Metabolic and Genome Engineering.” *Nature Biotechnology* 35 (1): 48–55.
Ghosh, Amit, Jerome Nilmeier, Daniel Weaver, Paul D. Adams, Jay D. Keasling, Aindrila Mukhopadhyay, Christopher J. Petzold, and Héctor García Martín. 2014. “A Peptide-Based Method for 13C Metabolic Flux Analysis in Microbial Communities.” *PLoS Computational Biology* 10 (9): e1003827.
Giaever, Guri, Angela M. Chu, Li Ni, Carla Connelly, Linda Riles, Steeve Véronneau, Sally Dow, et al. 2002. “Functional Profiling of the Saccharomyces Cerevisiae Genome.” *Nature* 418 (6896): 387–91.
Gibson, Daniel G., Gwynedd A. Benders, Cynthia Andrews-Pfannkoch, Evgeniya A. Denisova, Holly Baden-Tillson, Jayshree Zaveri, Timothy B. Stockwell, et al. 2008. “Complete Chemical Synthesis, Assembly, and Cloning of a Mycoplasma Genitalium Genome.” *Science* 319 (5867): 1215–20.
Gilbert, Luke A., Matthew H. Larson, Leonardo Morsut, Zairan Liu, Gloria A. Brar, Sandra E. Torres, Noam Stern-Ginossar, et al. 2013. “CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes.” *Cell* 154 (2): 442–51.
Gohl, Daryl M., Limor Freifeld, Marion Silies, Jennifer J. Hwa, Mark Horowitz, and Thomas R. Clandinin. 2014. “Large-Scale Mapping of Transposable Element Insertion Sites Using Digital Encoding of Sample Identity.” *Genetics* 196 (3): 615–23.
Gohl, Daryl M., Marion A. Silies, Xiaojing J. Gao, Sheetal Bhalerao, Francisco J. Luongo, Chun-Chieh Lin, Christopher J. Potter, and Thomas R. Clandinin. 2011. “A Versatile in Vivo System for Directed Dissection of Gene Expression Patterns.” *Nature Methods* 8 (3): 231–

37.
Goodall, Emily C. A., Ashley Robinson, Iain G. Johnston, Sara Jabbari, Keith A. Turner, Adam F. Cunningham, Peter A. Lund, Jeffrey A. Cole, and Ian R. Henderson. 2018. “The Essential Genome of Escherichia Coli K-12.” *mBio* 9 (1). https://doi.org/10.1128/mBio.02096-17.
Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. “Coming of Age: Ten Years of next-Generation Sequencing Technologies.” *Nature Reviews. Genetics* 17 (6): 333–51.
Goryshin, I. Y., J. Jendrisak, L. M. Hoffman, R. Meis, and W. S. Reznikoff. 2000. “Insertional Transposon Mutagenesis by Electroporation of Released Tn5 Transposition Complexes.” *Nature Biotechnology* 18 (1): 97–100.
Green, Brian, Christiane Bouchier, Cécile Fairhead, Nancy L. Craig, and Brendan P. Cormack. 2012. “Insertion Site Preference of Mu, Tn5, and Tn7 Transposons.” *Mobile DNA* 3 (1): 3.
Haddox, Hugh K., Adam S. Dingens, and Jesse D. Bloom. 2016. “Experimental Estimation of the Effects of All Amino-Acid Mutations to HIV’s Envelope Protein on Viral Replication in Cell Culture.” *PLoS Pathogens* 12 (12): e1006114.
Haldimann, A., and B. L. Wanner. 2001. “Conditional-Replication, Integration, Excision, and Retrieval Plasmid-Host Systems for Gene Structure-Function Studies of Bacteria.” *Journal of Bacteriology* 183 (21): 6384–93.
Haller, Gabe, David Alvarado, Kevin McCall, Robi D. Mitra, Matthew B. Dobbs, and Christina A. Gurnett. 2016. “Massively Parallel Single-Nucleotide Mutagenesis Using Reversibly Terminated Inosine.” *Nature Methods*, October. https://doi.org/10.1038/nmeth.4015.
Halweg-Edwards, Andrea L., Gur Pines, James D. Winkler, Assaf Pines, and Ryan T. Gill. 2016. “A Web Interface for Codon Compression.” *ACS Synthetic Biology* 5 (9): 1021–23.
Henikoff, S., and J. G. Henikoff. 1992. “Amino Acid Substitution Matrices from Protein Blocks.” *Proceedings of the National Academy of Sciences of the United States of America* 89 (22): 10915–19.
Hilton, Isaac B., Anthony M. D’Ippolito, Christopher M. Vockley, Pratiksha I. Thakore, Gregory E. Crawford, Timothy E. Reddy, and Charles A. Gersbach. 2015. “Epigenome Editing by a CRISPR-Cas9-Based Acetyltransferase Activates Genes from Promoters and Enhancers.” *Nature Biotechnology* 33 (5): 510–17.
Hug, Laura A., Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, et al. 2016. “A New View of the Tree of Life.” *Nature Microbiology* 1 (April): 16048.
Ideker, T., T. Galitski, and L. Hood. 2001. “A New Approach to Decoding Life: Systems Biology.” *Annual Review of Genomics and Human Genetics* 2: 343–72.
Islam, Saiful, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. 2011. “Characterization of the Single-Cell Transcriptional Landscape by Highly Multiplex RNA-Seq.” *Genome Research* 21 (7): 1160–67.
Jacobs, Michael A., Ashley Alwood, Iyarit Thaipisuttikul, David Spencer, Eric Haugen, Stephen Ernst, Oliver Will, et al. 2003. “Comprehensive Transposon Mutant Library of Pseudomonas Aeruginosa.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (24): 14339–44.
Jaffe, Mia, Gavin Sherlock, and Sasha F. Levy. 2017. “iSeq: A New Double-Barcode Method for Detecting Dynamic Genetic Interactions in Yeast.” *G3* 7 (1): 143–53.
Jaitin, Diego Adhemar, Assaf Weiner, Ido Yofe, David Lara-Astiaso, Hadas Keren-Shaul, Eyal David, Tomer Meir Salame, Amos Tanay, Alexander van Oudenaarden, and Ido Amit. 2016. “Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq.” *Cell* 167 (7): 1883–96.e15.
Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. 2000. “The Large-Scale Organization of Metabolic Networks.” *Nature* 407 (6804): 651–54.
Jiang, Wenyan, David Bikard, David Cox, Feng Zhang, and Luciano A. Marraffini. 2013. “RNA-Guided Editing of Bacterial Genomes Using CRISPR-Cas Systems.” *Nature Biotechnology* 31 (3): 233–39.

Jinek, Martin, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. 2012. “A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity.” *Science* 337 (6096): 816–21.

Jinek, Martin, Fuguo Jiang, David W. Taylor, Samuel H. Sternberg, Emine Kaya, Enbo Ma, Carolin Anders, et al. 2014. “Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation.” *Science* 343 (6176): 1247997.

Jin, Zhao, Sara C. Di Rienzi, Anders Janzon, Jeff J. Werner, Largus T. Angenent, Jeffrey L. Dangl, Douglas M. Fowler, and Ruth E. Ley. 2015. “Novel Rhizosphere Soil Alleles for the Enzyme 1-Aminocyclopropane-1-Carboxylate Deaminase Queried for Function with an In Vivo Competition Assay.” *Applied and Environmental Microbiology* 82 (4): 1050–59.

Ji, Weiyue, Derrick Lee, Eric Wong, Priyanka Dadlani, David Dinh, Verna Huang, Kendall Kearns, et al. 2014. “Specific Gene Repression by CRISPRi System Transferred through Bacterial Conjugation.” *ACS Synthetic Biology* 3 (12): 929–31.

Ji, Y., B. Zhang, S. F. Van, Horn, P. Warren, G. Woodnutt, M. K. Burnham, and M. Rosenberg. 2001. “Identification of Critical Staphylococcal Genes Using Conditional Phenotypes Generated by Antisense RNA.” *Science* 293 (5538): 2266–69.

Kaczmarczyk, Andreas, Julia A. Vorholt, and Anne Francez-Charlot. 2013. “Cumate-Inducible Gene Expression System for Sphingomonads and Other Alphaproteobacteria.” *Applied and Environmental Microbiology* 79 (21): 6795–6802.

Keseler, Ingrid M., Amanda Mackie, Alberto Santos-Zavaleta, Richard Billington, César Bonavides-Martínez, Ron Caspi, Carol Fulcher, et al. 2017. “The EcoCyc Database: Reflecting New Knowledge about Escherichia Coli K-12.” *Nucleic Acids Research* 45 (D1): D543–50.

Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. 2015. “HISAT: A Fast Spliced Aligner with Low Memory Requirements.” *Nature Methods* 12 (4): 357–60.

Kim, Seong Keun, Haseong Kim, Woo-Chan Ahn, Kwang-Hyun Park, Eui-Jeon Woo, Dae-Hee Lee, and Seung-Goo Lee. 2017. “Efficient Transcriptional Gene Repression by Type V-A CRISPR-Cpf1 from Eubacterium Eligens.” *ACS Synthetic Biology* 6 (7): 1273–82.

Kimura, Satoshi, Troy P. Hubbard, Brigid M. Davis, and Matthew K. Waldor. 2016. “The Nucleoid Binding Protein H-NS Biases Genome-Wide Transposon Insertion Landscapes.” *mBio* 7 (4). https://doi.org/10.1128/mBio.01351-16.

Kitagawa, Masanari, Takeshi Ara, Mohammad Arifuzzaman, Tomoko Ioka-Nakamichi, Eiji Inamoto, Hiromi Toyonaga, and Hirotada Mori. 2005. “Complete Set of ORF Clones of Escherichia Coli ASKA Library (a Complete Set of E. Coli K-12 ORF Archive): Unique Resources for Biological Research.” *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* 12 (5): 291–99.

Komor, Alexis C., Yongjoo B. Kim, Michael S. Packer, John A. Zuris, and David R. Liu. 2016. “Programmable Editing of a Target Base in Genomic DNA without Double-Stranded DNA Cleavage.” *Nature* 533 (7603): 420–24.

Konermann, Silvana, Mark D. Brigham, Alexandro E. Trevino, Julia Joung, Omar O. Abudayyeh, Clea Barcena, Patrick D. Hsu, et al. 2015. “Genome-Scale Transcriptional Activation by an Engineered CRISPR-Cas9 Complex.” *Nature* 517 (7536): 583–88.

Kong, Jun, Feng Wang, James D. Brenton, and David J. Adams. 2010. “Slingshot: A PiggyBac Based Transposon System for Tamoxifen-Inducible ‘Self-Inactivating’ Insertional Mutagenesis.” *Nucleic Acids Research* 38 (18): e173.

Krzywinski, Martin, Jacqueline Schein, Inanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J. Jones, and Marco A. Marra. 2009. “Circos: An Information Aesthetic for Comparative Genomics.” *Genome Research* 19 (9): 1639–45.

Landy, A., and W. Ross. 1977. “Viral Integration and Excision: Structure of the Lambda Att Sites.” *Science* 197 (4309): 1147–60.

Langridge, Gemma C., Minh-Duy Phan, Daniel J. Turner, Timothy T. Perkins, Leopold Parts, Jana Haase, Ian Charles, et al. 2009. “Simultaneous Assay of Every Salmonella Typhi Gene Using One Million Transposon Mutants.” *Genome Research* 19 (12): 2308–16.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, et al. 2007. “Clustal W and Clustal X Version 2.0.” *Bioinformatics* 23 (21): 2947–48.

Lazinski, David W., and Andrew Camilli. 2013. “Homopolymer Tail-Mediated Ligation PCR: A Streamlined and Highly Efficient Method for DNA Cloning and Library Construction.” *BioTechniques* 54 (1): 25–34.

Levy, Sasha F., Jamie R. Blundell, Sandeep Venkataram, Dmitri A. Petrov, Daniel S. Fisher, and Gavin Sherlock. 2015. “Quantitative Evolutionary Dynamics Using High-Resolution Lineage Tracking.” *Nature* 519 (7542): 181–86.

Li, Jingjing, Zineng Yuan, and Zhaolei Zhang. 2010. “The Cellular Robustness by Genetic Redundancy in Budding Yeast.” *PLoS Genetics* 6 (11): e1001187.

Lin-Goerke, J. L., D. J. Robbins, and J. D. Burczak. 1997. “PCR-Based Random Mutagenesis Using Manganese and Reduced dNTP Concentration.” *BioTechniques* 23 (3): 409–12.

Liu, X. Shawn, Hao Wu, Xiong Ji, Yonatan Stelzer, Xuebing Wu, Szymon Czauderna, Jian Shu, Daniel Dadon, Richard A. Young, and Rudolf Jaenisch. 2016. “Editing DNA Methylation in the Mammalian Genome.” *Cell* 167 (1): 233–47.e17.

Lorenzo, V. de, M. Herrero, U. Jakubzik, and K. N. Timmis. 1990. “Mini-Tn5 Transposon Derivatives for Insertion Mutagenesis, Promoter Probing, and Chromosomal Insertion of Cloned DNA in Gram-Negative Eubacteria.” *Journal of Bacteriology* 172 (11): 6568–72.

Ma, Jianzhu, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, and Trey Ideker. 2018. “Using Deep Learning to Model the Hierarchical Structure and Function of a Cell.” *Nature Methods*, March. https://doi.org/10.1038/nmeth.4627.

Malakhov, Michael P., Michael R. Mattern, Oxana A. Malakhova, Mark Drinker, Stephen D. Weeks, and Tauseef R. Butt. 2004. “SUMO Fusions and SUMO-Specific Protease for Efficient Expression and Purification of Proteins.” *Journal of Structural and Functional Genomics* 5 (1-2): 75–86.

Mandy, Dominic E., Joshua E. Goldford, Hong Yang, Doug K. Allen, and Igor G. L. Libourel. 2014. “Metabolic Flux Analysis Using ^{13}C Peptide Label Measurements.” *The Plant Journal: For Cell and Molecular Biology* 77 (3): 476–86.

Marblestone, Jeffrey G., Suzanne C. Edavettal, Yiting Lim, Peter Lim, Xun Zuo, and Tauseef R. Butt. 2006. “Comparison of SUMO Fusion Technology with Traditional Gene Fusion Systems: Enhanced Expression and Solubility with SUMO.” *Protein Science: A Publication of the Protein Society* 15 (1): 182–89.

Maxwell, K. L., A. K. Mittermaier, J. D. Forman-Kay, and A. R. Davidson. 1999. “A Simple in Vivo Assay for Increased Protein Solubility.” *Protein Science: A Publication of the Protein Society* 8 (9): 1908–11.

Meng, Jia, Gregory Kanzaki, Diane Meas, Christopher K. Lam, Heather Crummer, Justina Tain, and H. Howard Xu. 2012. “A Genome-Wide Inducible Phenotypic Screen Identifies Antisense RNA Constructs Silencing Escherichia Coli Essential Genes.” *FEMS Microbiology Letters* 329 (1): 45–53.

Mukherjee, Supratim, Rekha Seshadri, Neha J. Varghese, Emiley A. Eloe-Fadrosh, Jan P. Meier-Kolthoff, Markus Göker, R. Cameron Coates, et al. 2017. “1,003 Reference Genomes of Bacterial and Archaeal Isolates Expand Coverage of the Tree of Life.” *Nature Biotechnology* 35 (7): 676–83.

Nambi, Subhalaxmi, Jarukit E. Long, Bibhuti B. Mishra, Richard Baker, Kenan C. Murphy, Andrew J. Olive, Hien P. Nguyen, Scott A. Shaffer, and Christopher M. Sassetti. 2015. “The Oxidative Stress Network of Mycobacterium Tuberculosis Reveals Coordination between Radical Detoxification Systems.” *Cell Host & Microbe* 17 (6): 829–37.

Nordström, Kurt, and Santanu Dasgupta. 2006. “Copy-Number Control of the Escherichia Coli Chromosome: A Plasmidologist’s View.” *EMBO Reports* 7 (5): 484–89.

Oka, A., H. Sugisaki, and M. Takanami. 1981. “Nucleotide Sequence of the Kanamycin Resistance Transposon Tn903.” *Journal of Molecular Biology* 147 (2): 217–26.

Opijnen, Tim van, Kip L. Bodi, and Andrew Camilli. 2009. “Tn-Seq: High-Throughput Parallel

Sequencing for Fitness and Genetic Interaction Studies in Microorganisms." *Nature Methods* 6 (10): 767–72.
O'Toole, G. A., and R. Kolter. 1998. "Initiation of Biofilm Formation in Pseudomonas Fluorescens WCS365 Proceeds via Multiple, Convergent Signalling Pathways: A Genetic Analysis." *Molecular Microbiology* 28 (3): 449–61.
Otsuka, Yuta, Ai Muto, Rikiya Takeuchi, Chihiro Okada, Motokazu Ishikawa, Koichiro Nakamura, Natsuko Yamamoto, et al. 2015. "GenoBase: Comprehensive Resource Database of Escherichia Coli K-12." *Nucleic Acids Research* 43 (Database issue): D606–17.
Peters, Jason M., Alexandre Colavin, Handuo Shi, Tomasz L. Czarny, Matthew H. Larson, Spencer Wong, John S. Hawkins, et al. 2016. "A Comprehensive, CRISPR-Based Functional Analysis of Essential Genes in Bacteria." *Cell* 165 (6): 1493–1506.
Pires, Douglas E. V., David B. Ascher, and Tom L. Blundell. 2014. "mCSM: Predicting the Effects of Mutations in Proteins Using Graph-Based Signatures." *Bioinformatics* 30 (3): 335–42.
Popowska, Magdalena, and Agata Krawczyk-Balska. 2013. "Broad-Host-Range IncP-1 Plasmids and Their Resistance Potential." *Frontiers in Microbiology* 4 (March): 44.
Qi, Lei S., Matthew H. Larson, Luke A. Gilbert, Jennifer A. Doudna, Jonathan S. Weissman, Adam P. Arkin, and Wendell A. Lim. 2013. "Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression." *Cell* 152 (5): 1173–83.
Rakowski, Sheryl A., and Marcin Filutowicz. 2013. "Plasmid R6K Replication Control." *Plasmid* 69 (3): 231–42.
Rubin, E. J., B. J. Akerley, V. N. Novik, D. J. Lampe, R. N. Husson, and J. J. Mekalanos. 1999. "In Vivo Transposition of Mariner-Based Elements in Enteric Bacteria and Mycobacteria." *Proceedings of the National Academy of Sciences of the United States of America* 96 (4): 1645–50.
Sanjana, Neville E., Ophir Shalem, and Feng Zhang. 2014. "Improved Vectors and Genome-Wide Libraries for CRISPR Screening." *Nature Methods* 11 (8): 783–84.
Sarkisyan, Karen S., Dmitry A. Bolotin, Margarita V. Meer, Dinara R. Usmanova, Alexander S. Mishin, George V. Sharonov, Dmitry N. Ivankov, et al. 2016. "Local Fitness Landscape of the Green Fluorescent Protein." *Nature* 533 (7603): 397–401.
Schlecht, Ulrich, Zhimin Liu, Jamie R. Blundell, Robert P. St Onge, and Sasha F. Levy. 2017. "A Scalable Double-Barcode Sequencing Platform for Characterization of Dynamic Protein-Protein Interactions." *Nature Communications* 8 (May): 15586.
Schmidt, Thomas G. M., and Arne Skerra. 2007. "The Strep-Tag System for One-Step Purification and High-Affinity Detection or Capturing of Proteins." *Nature Protocols* 2 (6): 1528–35.
Schwartz, Jerrod J., Choli Lee, and Jay Shendure. 2012. "Accurate Gene Synthesis with Tag-Directed Retrieval of Sequence-Verified DNA Molecules." *Nature Methods* 9 (9): 913–15.
Sender, Ron, Shai Fuchs, and Ron Milo. 2016. "Revised Estimates for the Number of Human and Bacteria Cells in the Body." *PLoS Biology* 14 (8): e1002533.
Senecoff, J. F., P. J. Rossmeissl, and M. M. Cox. 1988. "DNA Recognition by the FLP Recombinase of the Yeast 2 Mu Plasmid. A Mutational Analysis of the FLP Binding Site." *Journal of Molecular Biology* 201 (2): 405–21.
Shafferman, A., and D. R. Helinski. 1983. "Structural Properties of the Beta Origin of Replication of Plasmid R6K." *The Journal of Biological Chemistry* 258 (7): 4083–90.
Shaikh, Afshan S., Yinjie J. Tang, Aindrila Mukhopadhyay, and Jay D. Keasling. 2008. "Isotopomer Distributions in Amino Acids from a Highly Expressed Protein as a Proxy for Those from Total Protein." *Analytical Chemistry* 80 (3): 886–90.
Shalem, Ophir, Neville E. Sanjana, Ella Hartenian, Xi Shi, David A. Scott, Tarjei Mikkelson, Dirk Heckl, et al. 2014. "Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells." *Science* 343 (6166): 84–87.
Shaner, Nathan C., Robert E. Campbell, Paul A. Steinbach, Ben N. G. Giepmans, Amy E. Palmer, and Roger Y. Tsien. 2004. "Improved Monomeric Red, Orange and Yellow Fluorescent Proteins Derived from Discosoma Sp. Red Fluorescent Protein." *Nature*

Biotechnology 22 (12): 1567–72.
Shendure, Jay, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss, and Robert H. Waterston. 2017. “DNA Sequencing at 40: Past, Present and Future.” *Nature* 550 (7676): 345–53.
Shen, John Paul, Dongxin Zhao, Roman Sasik, Jens Luebeck, Amanda Birmingham, Ana Bojorquez-Gomez, Katherine Licon, et al. 2017a. “Combinatorial CRISPR-Cas9 Screens for de Novo Mapping of Genetic Interactions.” *Nature Methods* 14 (6): 573–76.
Shuman, Stewart, and Michael S. Glickman. 2007. “Bacterial DNA Repair by Non-Homologous End Joining.” *Nature Reviews. Microbiology* 5 (11): 852–61.
Shu, Xiaokun, Nathan C. Shaner, Corinne A. Yarbrough, Roger Y. Tsien, and S. James Remington. 2006. “Novel Chromophores and Buried Charges Control Color in mFruits.” *Biochemistry* 45 (32): 9639–47.
Siegele, D. A., and J. C. Hu. 1997. “Gene Expression from Plasmids Containing the araBAD Promoter at Subsaturating Inducer Concentrations Represents Mixed Populations.” *Proceedings of the National Academy of Sciences of the United States of America* 94 (15): 8168–72.
Smith, Andrew M., Lawrence E. Heisler, Joseph Mellor, Fiona Kaper, Michael J. Thompson, Mark Chee, Frederick P. Roth, Guri Giaever, and Corey Nislow. 2009. “Quantitative Phenotyping via Deep Barcode Sequencing.” *Genome Research* 19 (10): 1836–42.
Stapleton, James A., Jeongwoon Kim, John P. Hamilton, Ming Wu, Luiz C. Irber, Rohan Maddamsetti, Bryan Briney, et al. 2016. “Haplotype-Phased Synthetic Long Reads from Short-Read Sequencing.” *PloS One* 11 (1): e0147229.
Starita, Lea M., Jonathan N. Pruneda, Russell S. Lo, Douglas M. Fowler, Helen J. Kim, Joseph B. Hiatt, Jay Shendure, Peter S. Brzovic, Stanley Fields, and Rachel E. Klevit. 2013. “Activity-Enhancing Mutations in an E3 Ubiquitin Ligase Identified by High-Throughput Mutagenesis.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (14): E1263–72.
Starita, Lea M., David L. Young, Muhtadi Islam, Jacob O. Kitzman, Justin Gullingsrud, Ronald J. Hause, Douglas M. Fowler, Jeffrey D. Parvin, Jay Shendure, and Stanley Fields. 2015. “Massively Parallel Functional Analysis of BRCA1 RING Domain Variants.” *Genetics* 200 (2): 413–22.
Sternberg, Samuel H., Sy Redding, Martin Jinek, Eric C. Greene, and Jennifer A. Doudna. 2014. “DNA Interrogation by the CRISPR RNA-Guided Endonuclease Cas9.” *Nature* 507 (7490): 62–67.
Szappanos, Balázs, Károly Kovács, Béla Szamecz, Frantisek Honti, Michael Costanzo, Anastasia Baryshnikova, Gabriel Gelius-Dietrich, et al. 2011. “An Integrated Approach to Characterize Genetic Interaction Networks in Yeast Metabolism.” *Nature Genetics* 43 (7): 656–62.
Takeuchi, Rikiya, Takeyuki Tamura, Toru Nakayashiki, Yuichirou Tanaka, Ai Muto, Barry L. Wanner, and Hirotada Mori. 2014. “Colony-Live--a High-Throughput Method for Measuring Microbial Colony Growth Kinetics--Reveals Diverse Growth Effects of Gene Knockouts in Escherichia Coli.” *BMC Microbiology* 14 (June): 171.
Taniguchi, Yuichi, Paul J. Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X. Sunney Xie. 2010. “Quantifying E. Coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells.” *Science* 329 (5991): 533–38.
Teixeira, Ana P., Sónia Sá Santos, Nuno Carinhas, Rui Oliveira, and Paula M. Alves. 2008. “Combining Metabolic Flux Analysis Tools and 13C NMR to Estimate Intracellular Fluxes of Cultured Astrocytes.” *Neurochemistry International* 52 (3): 478–86.
Tong, Amy Hin Yan, Guillaume Lesage, Gary D. Bader, Huiming Ding, Hong Xu, Xiaofeng Xin, James Young, et al. 2004. “Global Mapping of the Yeast Genetic Interaction Network.” *Science* 303 (5659): 808–13.
Tong, Yaojun, Pep Charusanti, Lixin Zhang, Tilmann Weber, and Sang Yup Lee. 2015. “CRISPR-Cas9 Based Engineering of Actinomycetal Genomes.” *ACS Synthetic Biology* 4 (9): 1020–29.

Typas, Athanasios, Robert J. Nichols, Deborah A. Siegele, Michael Shales, Sean R. Collins, Bentley Lim, Hannes Braberg, et al. 2008. “High-Throughput, Quantitative Analyses of Genetic Interactions in E. Coli.” *Nature Methods* 5 (9): 781–87.
Vaishampayan, Parag, Christine Moissl-Eichinger, Rüdiger Pukall, Peter Schumann, Cathrin Spröer, Angela Augustus, Anne Hayden Roberts, et al. 2013. “Description of Tersicoccus Phoenicis Gen. Nov., Sp. Nov. Isolated from Spacecraft Assembly Clean Room Environments.” *International Journal of Systematic and Evolutionary Microbiology* 63 (Pt 7): 2463–71.
Wall, M. A., M. Socolich, and R. Ranganathan. 2000. “The Structural Basis for Red Fluorescence in the Tetrameric GFP Homolog DsRed.” *Nature Structural Biology* 7 (12): 1133–38.
Wang, Harris H., Hwangbeom Kim, Le Cong, Jaehwan Jeong, Duhee Bang, and George M. Church. 2012. “Genome-Scale Promoter Engineering by Coselection MAGE.” *Nature Methods* 9 (6): 591–93.
Wang, Tim, Jenny J. Wei, David M. Sabatini, and Eric S. Lander. 2014. “Genetic Screens in Human Cells Using the CRISPR-Cas9 System.” *Science* 343 (6166): 80–84.
Wetmore, Kelly M., Morgan N. Price, Robert J. Waters, Jacob S. Lamson, Jennifer He, Cindi A. Hoover, Matthew J. Blow, et al. 2015. “Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons.” *mBio* 6 (3): e00306–15.
Wierzbicki, A., M. Kendall, K. Abremski, and R. Hoess. 1987. “A Mutational Analysis of the Bacteriophage P1 Recombinase Cre.” *Journal of Molecular Biology* 195 (4): 785–94.
Wilson, Thomas E., Leana M. Topper, and Phillip L. Palmbos. 2003. “Non-Homologous End-Joining: Bacteria Join the Chromosome Breakdance.” *Trends in Biochemical Sciences* 28 (2): 62–66.
Wiser, Michael J., and Richard E. Lenski. 2015. “A Comparison of Methods to Measure Fitness in Escherichia Coli.” *PloS One* 10 (5): e0126210.
Yachie, Nozomu, Evangelia Petsalaki, Joseph C. Mellor, Jochen Weile, Yves Jacob, Marta Verby, Sedide B. Ozturk, et al. 2016. “Pooled-Matrix Protein Interaction Screens Using Barcode Fusion Genetics.” *Molecular Systems Biology* 12 (4): 863.
Yamamoto, Natsuko, Kenji Nakahigashi, Tomoko Nakamichi, Mihoko Yoshino, Yuki Takai, Yae Touda, Akemi Furubayashi, et al. 2009. “Update on the Keio Collection of Escherichia Coli Single-Gene Deletion Mutants.” *Molecular Systems Biology* 5 (December): 335.
Zetsche, Bernd, Jonathan S. Gootenberg, Omar O. Abudayyeh, Ian M. Slaymaker, Kira S. Makarova, Patrick Essletzbichler, Sara E. Volz, et al. 2015. “Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System.” *Cell* 163 (3): 759–71.