# Multiplexed Genetic Perturbations of the Regulatory Network of E. coli.

Matthew Deyell

Thèse de doctorat de

L'Université Sorbonne Paris Cité

Préparée à l'Université Paris Diderot

**École doctorale Frontières du Vivant ED474**

ESPCI – PARISTECH

# Multiplexed Genetic Perturbations of the Regulatory Network of E. coli

Par Matthew DEYELL

Thèse de doctorat de biologie

Dirigée par Andrew GRIFFITHS et Philippe NGHE

Présentée et soutenue publiquement à Institut Pierre-Gilles de Gennes le 23/10/2018

Président du jury : Dillmann, Christine / Professeure / INRA, Université Paris-Sud Prénom

Rapporteurs : Tans, Sander / Professeur / AMOLF, TU Delft

Reynolds, Kimberly / Professeure Assistant / UT Southwestern

Examinateurs : Thieffry, Denis / Professeur / ENS

Sorre, Benoit / Maître de conférences / Université Paris Diderot

Directeur de thèse : Griffiths, Andrew / Professeur / ESPCI Paris

Co-directeurs de thèse : Nghe, Philippe / Maître de conférences / ESPCI Paris

**Titre : Perturbations Génétiques Multiplexées Du Réseau Régulateur De E. coli**

**Résumé :** Malgré les progrès réalisés dans le séquençage de l'ADN, nous n'avons pas encore compris comment le phénotype d'un organisme se rapporte au contenu de son génome. Cependant, il est devenu clair que l'impact des gènes dépend du contexte. La simple présence d'un gène dans un génome ne nous informe pas du moment où il est exprimé et des autres gènes qui y sont exprimés. Comprendre comment l'expression des gènes est régulée est un élément nécessaire pour comprendre comment les phénotypes émergent d'un génotype donné. Les facteurs de transcription, qui peuvent activer ou réprimer l'expression d'un gène, forment un réseau complexe d'interactions entre eux et leurs gènes ciblés. Ce réseau consiste en une hiérarchie de groupes de facteurs de transcription fortement liés, chacun lié à des processus cellulaires distincts. La structure de ce réseau de régulation transcriptionnelle est-elle significative pour la réponse transcriptionnelle d'une cellule? Ici, nous utilisons une protéine de liaison à l'ADN programmable appelée CRISPR (répétitions courtes palindromiques groupées régulièrement) pour perturber l'expression génique des régulateurs globaux au sein du réseau de régulation transcriptionnelle. Ces régulateurs mondiaux régulent de nombreux processus cellulaires distincts et ont de nombreuses cibles génétiques. Le système CRISPR nous permet de perturber ces régulateurs dans toutes les combinaisons possibles, y compris les perturbations d'ordre supérieur avec tous les régulateurs mondiaux potentiellement ciblés perturbés en même temps. Nous enregistrons ensuite à la fois le modèle d'expression du transciptome en utilisant le séquençage de l'ARN et l'adéquation de chaque souche. Nous trouvons que la structure du réseau de régulation augmente la dimensionnalité de la réponse transcriptionnelle plutôt que de la réduire. Cela se traduit par une épistasie importante au-delà des interactions par paires. Cela a des implications sur la façon dont ces réseaux évoluent. L'épistasie par paires que nous trouvons entre les facteurs de transcription globaux repose sur la présence ou l'absence d'autres perturbations. Cela implique que d'autres perturbations pourraient agir comme des mutations de potentialisation. Le nombre de voies d'évolution potentielles augmente avec les épistasies d'ordre élevé, même si cela ne nous dit rien sur la qualité de ces voies. Fait important, les répliques de cette thèse sont toujours en cours et les données présentées ici n'ont pas encore exclu les artefacts expérimentaux.

**Mots clefs : CRISPR, microfluidique, réseau transcriptionnel, Association génotype-phénotype, séquençage de l'ARN**

**Title : Multiplexed Genetic Perturbations of the Regulatory Network of E. coli**

**Abstract :** Despite advances in DNA sequencing, we have yet to understand how an organism's phenotype relates to the contents of their genome. However it has become clear that the impact of genes are context dependant. The mere presence of a gene within a genome does not inform us of when it is expressed, and which other genes are expressed along with it. Understanding how gene expression is regulated is a necessary piece of understanding how phenotypes emerge from a given genotype. Transcription factors, which can activate or repress the expression of a gene, form a complex network of interactions between themselves and their targeted genes. This network consists of a hierarchy of groups of strongly connected transcription factors, each relating to distinct cellular processes. Is the structure of this transcriptional regulatory network significant to the transcriptional response of a cell? Here we use a programmable DNA binding protein called CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) to perturb gene expression of global regulators within the transcriptional regulatory network. These global regulators are regulating many distinct cellular processes and have many genetic targets. The CRISPR system allows us to perturb these regulators in all possible combinations, including higher order perturbations with potentially all targeted global regulators perturbed at the same time. We then record both the expression pattern of the transciptome using RNA sequencing, and the fitness of each strain. We find that the structure of the regulatory network increases the dimensionality of the transcriptional response rather than reducing it. This results in significant high order epistasis beyond pair-wise interactions. This has implications for how these networks evolve. The pair-wise epistasis we find between global transcription factors rely on the presence or absence of other perturbations. This implies that other perturbations could act as potentiating mutations. The number of potential evolutionary paths increases with high order epistasis, although this alone tells us nothing about the quality of those paths. Importantly, the replicates for this thesis are still on-going and the data presented here has not yet excluded experimental artefacts.

**Keywords : CRISPR, Microfluidics, Transcriptional Network, Genotype to Phenotype Mapping, single-cell RNA sequencing**

# Multiplexed Genetic Perturbations of the Regulatory Network of E. coli

Doctoral Thesis of Matthew Deyell

Supervisors: Andrew Griffiths and Philippe Nghe

ESPCI – PARISTECH 10 Rue Vauquelin

# Table of Contents

# *Acknowledgements*

I would like to thank:

**Andrew Griffiths** and **Philippe Nghe** for advising me during my thesis and giving me the opportunity to work with them at ESPCI-Paris.

**Kimberley Reynolds** and **Sander Tans** for agreeing to be rapporteurs for my thesis.

**Christine Dillmann**, **Denis Thieffry**, and **Benoit Sorre** for being members of my thesis jury.

**David Bikard** and **Pascal Hersen** for being my doctoral tutors and helping me with advice on CRISPR and genetic control respectively

**Angga Perima** for assistance with droplet-based microfluidics, fruitful discussions, and friendship.

**Claire Seydoux**, **Gabrielle Schanne**, **Marie Baumont**, **Milan Lacassin** and **Marine Lombard** for their hard work during their internships.

**Isabelle Borsenberger** and **Helene Dodier** for all their help and for keeping the lab running.

**Sean Kennedy**, **Jean-Yves Coppee**, and **Odile Sismeiro** for their collaboration with single cell RNA sequencing.

**Sophie Foulon** and **Baptiste Saudemont** for help with single cell RNA sequencing and barcoded hydrogel beads.

**Clément Nizak**, **Ariel Lindner**, **Marcel Reichen** and **Jake Wintermute** for guidance and advice.

**Sandeep Ameta**, **Roberta Menafra**, **Marco Ribezzi**, **Pablo Ibanez**, and **Heng Lu** for help with molecular biology and microfluidics.

**Simon Arsène**, **Alex Blokhuis** and **Kevin Grosselin** for their help with bioinformatics.

**Dany Chauvin**, **Stéphane Chiron**, **Raphaël Doineau**, **Andrea Flamm**, **Andrea Franconi**, **Adeline Pichard Kostuch**, **Marina Theodorou**, **Robin Tranchant**, **Anton Zadorin**, **Carlos Castrillon**, **Thomas Duboys**, **Guillaume Mottet**, **Amandine Trouchet**, **Gabrielle Woronoff**, **Susan Hassan**, **Guillaume Mottet**, **Timothy Kirk**, **Judith Boldt** and **Guillaume Villian** for their friendship and advice.

**Elodie Kaslikowski** for all her help and assistance.

**David Tareste** and the rest of the CRI for allowing me to be a part of the FdV doctoral school.

**Patrick Tabeling**, **Perrine Franquet**, and the rest of IPGG for funding my PhD and providing support and advice with microfluidics.

**Sebastian Jaramillo Riveri**, **Aude Bernheim**, **Vincent Libis** and **Marguerite Benony** for their friendship and support.

**Chelsie Loveder** for her constant support and encouragement.

# 1 Introduction

One of the outstanding problems in biology is predicting phenotype from a given genotype, and how changes to genotype will influence phenotype. Undoubtedly a landmark moment, the human genome project promised to dramatically accelerate diagnosis, prevention and treatment of disease. This would not be limited to single-gene disorders but also complex diseases such as heart disease, schizophrenia and cancer [1]. Upon completion of the human genome, a blueprint for the future of genomics was proposed as a building with the human genome project as the foundation [2]. The first floor of this blueprint was genomics to biology, and set out 3 grand challenges for us to accomplish next. These included a complete catalogue of all the components encoded in the human genome, how genome-encoded components function together to give rise to functions on the cellular and organism scales, and understanding how genomes change to take on new functional roles. The advent of deep sequencing techniques has resulted in an explosion of genomic information, yet completion of these three challenges remains elusive. Identifying all of the components in the human genome is difficult. In 2018, an additional 348 human transcription factors were identified, and the current list is likely still incomplete [3]. Cancer genomics has highlighted the challenges in understanding how the components we do know interact. The 'one gene, one function, one disease' model does not match observations that different mutations in the same gene result in a variety of different phenotypes [4]. Finally, rather than understanding how genomes evolve to produce new functions, new information from genetics is forcing us to reconsider fundamental assumptions in evolution. The gene has long been considered the unit of inheritance [5], however the difficulty in linking them to the phenotypes needed for natural selection has called this into question. It has become apparent that the expression profile of genes, that is to say how and when they are expressed, as well as post-transcriptional events are just as important as the genetic sequence in determining function.

With such complexity, it is useful to turn to our trusty model organism *Escherichia coli*. *E. coli* is the most well studied organism in the world [6]. It has the most exhaustively annotated genome at only 4.6 MB compared to the 3,234.8 MB in the human genome but it also shares many of the same challenges. We still do not know how genetic elements interact to form complex phenotypes. For example, the mere presence of a virulence factor is not predictive of a virulent phenotype in clinical settings. The pathogenicity of a bacterium is conferred by both the virulence factors it possesses and the immune status of the host. Crucially, this interplay between host and pathogen depends on how and when these virulence factors are expressed [7]. As a result, the same virulence factors located in different bacterial genomes can result in different phenotypes. This phenotypic heterogeneity arising from the same genetic sequence is even more pronounced in the case of antibiotic persistence. In this example, within a bacterial population of clonal, genetically identical cells there is a phenotypically distinct subpopulation, which survive antibiotic treatment despite lacking any antibiotic resistance. This phenomenon is characterised by a classic biphasic killing curve when treated with antibiotics, in which a susceptible sub-population displays rapid death, and a persistent population has a much flatter slower death rate [8]. Persistence is a phenotypic switch that does not involve genotypic change, and progeny from persister cells will retain the same killing dynamics as the original population. How these multiple phenotypes arise in a genetically clonal population remains unknown.

Deep sequencing has allowed us to measure the entire transcriptome of a population of cells, which is all the mRNA expressed at a given time. This allows for a 'Top-Down' systems biology approach to understanding how gene expression is controlled at a global scale. The difficulty remains that gene expression profiles are the result of highly entangled interactions between many genes [9], statistical inference from large scale "omics" screens are at most validated from internal consistency [10], and genetic and protein interaction networks have only limited overlap in micro-organisms [11]. Part of these problems arise from measurement errors and batch effects biasing the resulting data [12], while experimental design may explain the remaining difficulties. Current experimental data is generally limited to single or pair-wise genetic perturbations, with poorly understood predictive power [13]. Most screens also consist of measuring only a few parameters for many mutants [14] or the full transcriptome for only a few mutants [15].

Recent technological advances can address these issues. A bacterial phage defense system, CRISPR (**C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats) allows for the programmable repression of gene expression in bacteria [16]. This allows for a simple method to quickly create a library of thousands of genetically perturbed strains, which can be multiplexed to knock-down multiple genes at a time. This has already been paired with another recent development, massively parallelized single-cell RNA-seq [17] [18]. By performing cDNA synthesis on RNA from single cells in nanolitre droplets, many thousands of RNA-seq experiments can be done at once. When used on a population of cells each containing a unique CRISPR perturbation, the full transcriptome of thousands of perturbations can be recorded in a single experiment [19] [20].

Here we develop and deploy these strategies in the context of the regulatory network of *E. coli*. Complex interdependences between genes have been the subject of diverging conclusions. On one hand, an elementary superimposition principle has been proposed for gene responses [21], down to the point that *E. coli* transcriptional programs may reduce to an interpolation between growth and starvation [22]. On the other hand, the complexity of regulatory systems and the cooperative nature of biological interactions suggest a much richer phenomenology [23], and consistently, network control theory shows that a high number of genes need to be controlled to drive the cell as a whole in a desired state [24].

Here we pose the following question: is there a relationship between the connections in transcriptional regulatory network and the response of the system to perturbations? While this seems straight-forward, we lack the parameters necessary to make an explicit model. We can however measure the dimensionality of the transcriptional response. This approach allows us to assess the connectivity in a way that does not depend on non-linearities such as expression saturation, cooperatively, signal integration, or the sign of the interactions. These unknown parameters will affect the shape of the response, but will not affect its dimensionality. This would allow us to determine if transcription factors which co-regulate each other have a coupled transcriptional response. It also informs us if we are able to reach more than just two transcriptional responses corresponding to growth and starvation, or if transcriptional programs are more nuanced. By incorporating these transcriptional programs with phenotypic data on our strains, we can then infer how variations in our transcription patterns correspond with variations in our phenotypes.

In the first chapter, we begin with a review of the transcriptional network of *E. coli*. We then discuss the global transcriptional regulators and their known roles in the cell, which we will later perturb to observe the transcriptional response. We present mathematical and computation models for analysing

the transcriptional regulatory network. These techniques help us with the interpretation of our data. We then present a review of CRISPR-Cas techniques for genetic control, to be published in Seminars in Cell and Development Biology this winter. This describes the various techniques with which CRISPR can be used for genetic perturbations.

Our second chapter involves further exploring the epistasis between transcription factors in different regulatory clusters. We found that (1) each global regulator is strongly connected to a cluster of regulators that mutually regulate each other, (2) these clusters have hierarchical relations associated with physiological tasks (oxidation, starvation…) and (3) all regulators are downstream combinations of these few clusters. Hierarchy in transcriptional cascades has been shown to cause sign epistasis [25], which can restrict the evolutionary paths that a cell may take. We quantify two fitness measurements for single and pairwise perturbations for transcription factors in different clusters in the regulatory network to determine if there is sign epistasis between regulators and if so, is it related to the structure of the network. We observe that similar to the effect of variable environments on sign epistasis [26], variable selection pressure (from multiple fitness metrics) may provide additional evolutionary paths with which to escape reciprocal sign epistasis.

Our third chapter involves using CRISPR-Cas to perturb the global regulatory transcription factors. We first quantify the effectiveness of the CRISPR-Cas system to knock-down gene expression using single perturbations. We then discuss a method for creating multiplexed perturbations, of libraries with a size of 2^N. The perturbation set consists of 32 different combinations of 5 knock-down targets, covering the genes: *arcA, crp, fis, fnr, hns*. These global regulators known to interact together from specific gene studies [27], but we lack general rules for the logic of this regulation. Existing studies have focused either on a single master regulator, some on a pair [28], or have analyzed aggregated microarray datasets [29]. We therefore create a systematic data set around a set of conditions, consisting of the fitness and transcriptional profiles for all 32 strains, in three growth media. This allows us to determine the higher order epistasis between global transcriptional regulators, the dimensionality of the genetic response, the different genetic programs and the regulatory logic for each of those programs.

Our fourth chapter explains how we can use microfluidics and molecular barcoding to measure epistasis between multiple environments in a high-throughput manner. Specifically, we check for epistasis between antibiotic drug combinations by associating each environment (antibiotic concentration) with a unique DNA barcode. These barcodes are then covalently linked in droplets to associate which combination of antibiotics was contained in each droplet. After selection of droplets that display bacterial growth, the drug combinations that led to that phenotype can be recovered by sequencing barcode pairs from each population of droplets. The epistasis between drug combinations can be determined by the shape of the threshold between combinations where growth occurs and where it does not. This chapter is Matthew Deyell's contribution to the thesis of Angga Perima entitled Combinatorial Antibiotic Screening Using Droplet Based Microfluidics.

Finally, our fifth chapter demonstrates the adaption of droplet-based single cell RNA-seq approaches for use with bacteria. Each droplet contains a clonal population of unique DNA bar-codes carried on a hydrogel bead. These bar-codes are appended to the genetic material of interest during RT-PCR inside the droplet. Genetic perturbations can be identified *a posteriori* by sequencing of DNA codes contained within the perturbation vector of the cell and gene expression can be quantified by RNA-seq. After all the genetic material present in a same droplet has been attached to a specific bar-code, one can

break the emulsion and pool the material for a single Next Generation Sequencing (NGS) run. The perturbation and corresponding gene expression measured in each droplet are then retrieved by bioinformatics sorting of the bar-codes. When coupled to the CRISPR perturbation strategy in Chapter 2, this would allow for RNA-sequencing of a large perturbation library, which can easily be expanded.

Taken together, these results indicate that the structure of the regulatory network increases the dimensionality of the gene expression response rather than reduce it. This may have implications in how transcriptional regulatory networks evolve, as increasing the dimensions may provide indirect paths for evolution [30]. However, first we must replicate our RNA-sequencing data and do so in multiple growth media. Also, if we wish to make a mechanistic claim, we should demonstrate it in a minimal model as well. We expect replicates and experiments for chapters 2, 3, and 4 to be completed soon, however these replicates are needed to exclude potential experimental artifacts. Regardless, these initial results are consistent with the many observations that changes in global regulation through global transcriptional regulators are some of the earliest and most common mutations when cells adapt to a new environment [31] [32] [33] [34] [35].

## 1.1 *E. coli* Transcriptional Regulatory Network

Since François Jacob and Jacques Monod described regulation of the lac operon [36], the standard model for transcriptional control in bacteria has been that special DNA binding proteins called transcription factors control when genes are expressed. These transcription factors are able to respond to the environment of the cell and adjust which genes are turned on as a result, to produce the necessary proteins required for that environment. However, transcription factors don't just control the expression of genes linked to a specific cellular function, but can also control the expression of other transcription factors. The interactions between these transcription factors form the transcriptional regulatory network of *E. coli*.



*Figure 1: Transcriptional regulatory network in E. coli in 2004.Yellow ovals demonstrate regulated genes, green ovals represent transcription factors, and blue ovals are global regulators. Lines represent activation (green), repression (red), or both (dark blue). [37]*

*By representing the system of transcription factors regulating other transcription factors as a network or graph, we can analyse the characteristics of the system. These include identifying network motifs such as feed forward loops, single input modules, and dense overlapping regulons [38]. A feed forward loop can cause transcription to only occur to a persistent stimuli. Single input modules are associated with protein complexes, where all components must be expressed together for the protein complex to function. These two motifs are usually connected to the output of the third motif, dense overlapping regulons, which represents the core regulatory system dominated by global transcription factors. Many of these network motifs can be combined into modules, which could drive a specific cellular process [32]. However the regulatory network is highly interconnected, with few truly distinct modules identified. Originally, the transcriptional regulatory network of E. coli was reported to exist in a hierarchical*

*structure, with no strongly connected components [39]. These strongly connected components are groups of transcription factors, in which each member can regulate the expression of every other transcription factor within the group. This resulted in a transcriptional network in which the global transcriptional regulators occur on the top level, with other transcription factors below them, and finally operons on the bottom level (*

Figure 1) [37].

**élément sous droit, diffusion non autorisée**

*Figure 2: Functional regulatory modules in E. coli transcriptional regulatory network. The size of each rectangle is proportional to the number of genes regulated by each transcription factor, although overlap is not shown [40].*

This model has recently been updated to describe ten functional regulatory modules [40]. These regulatory modules correspond to distinct cellular processes (

élément sous droit, diffusion non autorisée

Figure 2). This has improved the prediction of differentially expressed genes from distinct perturbations, (23% average compared to 15%) although the results make it clear that there are additional cellular processes that influence gene expression beyond just the transcriptional regulatory network.

Our ability to predict the controllability of the regulatory network of the cell on gene expression is confounded by its relation to another hierarchical organization in the form of chromosomal structures which has been shown to be implicated in cell cycle coordination, transition to virulence, transitions between growth phases, and stress, thermal and osmotic responses [41]. Genes recently acquired by horizontal transfer are exclusively regulated by nucleoid associated proteins (NAPs), furthermore suggesting that this is the default mode of regulation. More generally in *E. coli* most genes are not specifically regulated by transcription factors (TF) and many of the master regulators are NAPs [13]. The latter are known to bend, wrap, bridge or coat DNA to modulate transcription at all scales of the nucleoid [21]. However, direct evidence of the impact of nucleoid organization on gene expression remains limited to a few genomic locations and we lack a consistent picture of the genome-wide scale at which this regulation operates.

One possible explanation for how nucleoid structures contribute to gene regulation is via facilitated co-transcription, the transcription of an operon is facilitated by the transcription of the operon located immediately upstream. This includes transcriptional read through which is known to be a major source of transcripts in bacteria. Additionally the torsional stress induced by transcribing RNA polymerases is known to enhance or repress the transcription of nearby genes. Due to their role in defining supercoiling domains, this is another mechanism in which global transcription factors may regulate gene expression in a sub-optimal manner. It is also possible for transcription factors to extend their influence from effects observed in synteny segments [42]; that is the conservation of relative distances between orthologous genes in different species is indicative of co-expression of those genes which is not explained purely by operons or common sigma factors or transcription factors. Genes which have no transcription factor of their own are significantly more likely to be co-expressed when they have exactly the same transcription factors acting on other genes in their synteny segment, indicating that a transcription factor can indirectly regulate the transcription of genes structurally associated with genes that are directly regulated.

The connection between structure and function remains elusive, as exemplified by the low overlap between genetic and protein interaction networks obtained in micro-organisms [9]. Despite sophisticated statistical methods, it remains highly challenging to infer functional relations from currently available genetic screens. Beyond the problem of measurement errors, large screens may actually suffer from methodological limitations that are not primarily related to inference methods: (i) there is little if no back and forth validations between measurements and models; (ii) perturbations are in large majority limited to pair-wise interactions, which predictive power is poorly understood in the context of cellular networks; (iii) most screens consist of measuring one or a few parameters for many strains [10], or the other way round, full gene expression for a handful of strains [14]. Increasing the measurement throughput and the level of multiplexing of genetic perturbations and read-outs would fundamentally modify the view of a problem of such complexity.

## 1.2 Global Transcriptional Regulators

The degree distribution for transcription factors in the regulatory network of *E. coli* is not uniform. There are hubs in the network, consisting of transcription factors with a large number of targets. Seven transcription factors are sufficient to directly modulate 51% of the genes in *E. coli* [37]. These genes are CRP, FNR, IHF, Fis, ArcA, NarL, and Lrp. Additionally, 49% of genes are regulated by multiple transcription factors, with global regulators working alongside more specific regulators or other global regulators. The high level of cooperativity between global regulators grants the cell flexibility to tune transitions between groups of co-regulated genes between conditions [37]. Here we summarize the global transcription factors as defined by Martinez-Antonio and Collado-Vides, based on a previous definition by Gottesman [43]. Global regulators are defined by their pleiotropic phenotype and ability to regulate operons belonging to multiple different metabolic pathways. They must also not be part of essential cellular machinery. This list includes the following 7 transcription factors: CRP, FNR, ArcA, Fis, H-NS, IHF, and LRP.

***CRP***

The cAMP receptor protein (CRP) is a transcriptional dual regulator that controls over 521 genes in *E. coli*, many of which are involved in the catabolism of secondary carbon sources. CRP is able to sense the energetic status of the cell by cAMP levels [37]. CRP is also involved in processes such as osmoregulation, stringent response, biofilm formation, virulence, nitrogen assimilation, iron uptake, competence, and multi-drug resistance. CRP is positively and negatively auto regulated and repressed by Fis. CRP is activated by binding of cAMP [44].

CRP is a member of the CRP-FNR superfamily of transcription factors (Figure 3). It is a homodimer and consists of two domains attached by a hinge. CRP binds to a 22-bp symmetrical site and induces a severe bend of approximately 80 degrees in DNA. Two regions of CRP are known to interact with RNA polymerase. Promoters activated by cAMP-CRP are grouped into three classes. Classes I and II have a single binding site for cAMP-CRP located upstream (Class I) or overlapping (Class II) the RNA polymerase binding site. Class III promoters require multiple activator molecules, either as multiple copies of CRP or in synergy with other transcription factors. CRP can act as a repressor by promoter exclusion, exclusion of another activator, in an antiactivation mechanism with a repressor or by hindering promoter clearance.

Due to autoregulation, CRP levels are highly correlated with cAMP levels in the cell. In the absence of a rapidly metabolized carbon source such as glucose, glucose-specific enzyme IIA becomes phosphorylated and activates adenylate cyclase resulting in elevated levels of cAMP. In the presence of glucose, enzyme IIA is dephosphorylated and cAMP levels drop [44].

***FNR***

FNR (fumarate and nitrate reductase) mediates the transition from aerobic to anaerobic growth. It activates genes involved in anaerobic metabolism and represses genes involved in aerobic metabolism. It is also involved in functions such as acid resistance, chemotaxis, cell structure, and molecular biosynthesis. The cellular concentration of FNR is similar under both anaerobic and aerobic growth however its activity is regulated directly by oxygen. FNR requires a 4Fe-4S cluster for dimerization and activation of the transcription factor. In the presence of oxygen this cluster is oxidized into a 2Fe-2S cluster and the FNR dimer disassembles and the monomer is able to be degraded by ClpXP protease. Activated FNR is able to bind the consensus sequence TTGATNNNNATCAA. This exposes three activating regions on FNR which are able to interact with RNA polymerase [44]. Under anaerobic growth conditions FNR is

negatively autoregulated. It is also negatively regulated by phosphorylated ArcA, and Fur while being positively regulated by IHF.

**élément sous droit, diffusion non autorisée**

*Figure 3: Transcription factors in their evolutionary families based on PFAM, CDD, and superfamily annotations. CRP and FNR exist in the same family (right). ArcA exists in the OmpR family (left). Fis exists in its own family (right). LRP is in the AsnC family (upper right). IHF and HNS are not mapped [45].*

***ArcA***

ArcA (aerobic respiration control) is a dual transcriptional regulator for anoxic redox control. It is primarily a negative transcriptional regulator under anaerobic conditions. ArcA represses operons involved in respiratory metabolism as well as rpoS. ArcA activates operons involved in fermentative metabolism. There is a large overlap between the Arc and FNR regulatory systems. It is suggested that the most significant fole of ArcA is under microaerobic conditions, while that of FNR is under more strictly anaerobic conditions [44].

ArcA is activated by phosphorylation by the cognate sensor kinase ArcB under anaerobic conditions. ArcB is simulated by effectors such as lactate, pyruvate and acetate, while under aerobic conditions it is inhibited by quinone electron carriers. ArcA is a member of the OmpR/PhoB subfamily of response regulators. ArcA represses transcription by directly binding to the promoter or by binding to sites overlapping an activator-binding site [44].

***Fis***

Fis (factor for inversion stimulation) is involved in the organization and maintenance of the nucleoid structure through direct DNA binding and modulating gyrase and topoisomerase I production, as well as regulation of other proteins that modulate nucleoid structure such as CRP, HNS and HU. Fis directly modulates several cellular processes such as transcription, chromosomal replication, DNA inversion, phage integration/excision, and DNA transposition. It is involved in the regulation of genes involved in translation (rRNA and tRNA genes), virulence, biofilm formation, energy metabolism, stress response, central intermediary metabolism, amino acid biosynthesis, transport, cell structure, carbon compound metabolism, amino acid metabolism, nucleotide metabolism, motility, and chemotaxis [44].

Fis is one of the largest components of the nucleoid. Fis binds to 894 regions in the genome resulting in two Fis sites per supercoiling domain. Under optimal growth conditions Fis is the dominant DNA binding protein in the cell. Fis can vary from up to 60, 000 copies per cell in log phase to less than 100 in stationary phase. Fis bends DNA at an angle between 40 and 90 degrees. This bending promotes DNA compaction and stabilizes DNA looping to regulate transcription [44].

Fis is regulated by several processes at different levels of control. Transcriptionally, Fis is auto-regulated and induced by high supercoiling. Transcription is regulated by the availability of CTP, which has its highest concentration during log phase. DksA increases the inhibitory effects of ppGpp, decreasing the half-life of the RNA-polymerase complex and increasing sensitivity to CTP. Fis binds to a degenerated consensus sequence of 15 bp with only four highly conserved nucleotides, a G in the first position, a pyrimidine in the $5^{th}$ position, a purine in the $11^{th}$ position, and a C in the $15^{th}$ position. The central region commonly presents as an AT-rich sequence [44].

***H-NS***

H-NS (Histone-like nucleoid structuring protein) is a nucleoid associated protein that is capable of condensing and supercoiling DNA. It also acts as a silencer of genes with high AT rich content and as such has a strong preference for horizontally acquired genes. It functions almost exclusively as a transcriptional repressor. H-NS induces severe bends in DNA and is able to form DNA-HNS-DNA bridges forming multimers. DNA binding is similar to StpA and these proteins have similar functions. StpA seems to be a back up for H-NS and can complement an H-NS mutant when highly expressed from a plasmid. H-NS

interacts with StpA to prevent degradation of StpA in a Lon protease dependant manner. H-NS may also form heterotrimeric complexes with Hha and YdgT. It is proposed that Hha enhances the oligomerization of H-NS/StpA. H-NS is capable of controlling its own expression and plays an important role in the cellular response to environmental changes and stress [44].

***IHF***

Integration host factor (IHF) is a global regulatory protein that maintains DNA architecture. It binds and bends DNA at specific sites and plays a role in DNA supercoiling and DNA duplex destabilization. IHF acts mostly as an accessory factor, stabilizing the nucleoprotein complex. In the case of lambda phage integration, IHF bends DNA to facilitate binding of the integrase. It also stimulates $\sigma^{54}$ dependant promoters by facilitating the DNA loop between the upstream activator and the $\sigma^{54}$ holoenzyme. Similar to HU, IHF plays a role in DNA condensation. This is done by binding of low affinity sites and introducing sharp bends of approximately 160 degrees to promote the formation of rod like structures in chromosomal DNA [44].

IHF is a heterodimer consisting of two subunits, IhfA and IhfB. It binds to a 40bp region containing a 13bp consensus sequence with an AT rich element upstream. Because IHF makes no contacts with the major groove of DNA and only a few contacts with the minor groove, it is believed that specificity is a result of the sequence specific structural characteristics of the DNA [44].

***LRP***

LRP (Leucine-responsive regulatory protein) is a dual transcriptional regulator for genes involved in amino acid biosynthesis and catabolism, nutrient transport, pili synthesis and carbon metabolism. It is able to sense the nutritional state of the cell through the leucine concentration and adjusts the cellular metabolism accordingly [37]. LRP may also play a role in dynamic DNA packaging. LRP can act as either an activator or repressor and the binding of leucine can affect these activities either positively, negatively or not at all. It is believed that LRP positively regulates genes that function during starvation and negatively regulates genes functional during feasting. LRP-regulated genes commonly contain multiple binding sites for LRP with low sequence specificity. The consensus sequence is a central AT rich sequence with flanking CAG/CTG triplets. The binding of Leucine results in a decrease in binding affinity but an increase in cooperatively to multiple binding sites [44].

LRP forms a mixture of octamers and hexadecamers. In the presence of leucine the octamer configuration is favored. It is believed that the switching between these two forms is how leucine modulates LRP binding. In the octamer form LRP forms a ring structure with DNA wrapping around the octamer in a nucleosome-like structure [44].

## 1.3 Mathematical and Computation Methods for modeling Regulatory Networks

Explanatory and predictive mathematical models are useful for understanding how transcriptional regulatory networks control gene expression patterns and encode different genetic programs [46]. Various approaches have been used to attempt to explain gene expression profiles, starting from either a structural view of the regulatory network, or from gene expression data in the form of micro-array data, or RNA-sequencing. Here we discuss some of these approaches, including structural controllability, pareto optimality, principle component analysis (PCA), and logical modelling. Mathematical and computational approaches are essential because transcriptional regulatory networks are complex systems; they are composed of a large number of non-linear and difficult to predict components [47].

***Structural controllability***

Although control theory is a highly developed branch of mathematics and engineering with many diverse applications, questions regarding the controllability of natural complex systems have resisted advances. Control theory states that a dynamic system is controllable if suitable inputs can drive the system from any initial state to any desired final state within a finite time [48] [49] [50]. Two factors contribute to the difficulty in the application of controllability to natural systems; the system's architecture which can be represented by a network of interacting components and the rules that govern the time dependant interactions between components. With natural systems we often lack information in both the architecture and rules and thus progress has mainly been possible in systems which both are fairly well mapped such as synchronized networks and small biological circuits [51]. Despite the nonlinear processes that drive most real systems, theoretical approaches to the controllability of real networks begin with the canonical linear, time invariant dynamics of control theory [24]:

$$\frac{d\mathbf{x}(t)}{dt} \sim A\mathbf{x}(t) + B\mathbf{u}(t)$$

In which, in greatly simplified terms, the change of the state of the system x depends on the first term which represents the initial state of the system and the second term which represents the input from the controller. The N x N matrix A describes the connections between the nodes in the system while B is an N x M ($M \leq N$) input matrix that identifies nodes controlled by an outside controller. The system described above can only be driven from any initial state to any desired final state within a finite time if the N x NM controllability matrix $C = (B, AB, A^2B, \ldots, A^{N-1}B)$ has full rank [24] such that rank(C) = N. This is referred to as Kalman's controllability rank condition.

To control a system we first must identify the 'driver nodes' which allow full control over the network. The minimum number of nodes required to maintain full control of the network is determined by the maximum set of links that do not share start or end nodes. We can gain full control over a network only if we can directly control each node that has no links pointing at it and there are directed paths from the input signals to all other nodes [24]. Given that the purpose of the gene regulatory network is to control the dynamics of cellular processes, it could reasonably be expected that it would evolve to be structurally efficient from a control perspective requiring only a small number of driver nodes. However analysis of the regulatory networks of yeast and *E. coli* indicate that for complete control of the system between 75-96% of nodes must be driver nodes [24]. Given the importance that regulatory hubs have in

genetic networks in regard to maintaining structural integrity against failures, spreading phenomena, and in synchronization we expect control of hubs to be essential to the control of the network. However, the fraction of driver nodes is significantly higher among nodes with low degrees of connectivity than hubs, indicating that hubs tend not to be driver nodes in natural systems [24]. Calculating the controllability matrix C when only the global regulators are perturbed generates a rank of only 24 (N=956) for the transcription factor - operon network and only 11 (N=174) for the transcription factor – transcription factor network as per interactions found on RegulonDB [45].

The approach used above has been criticized for over-estimating the driver nodes required to give complete control over a system because each node is assumed to have an infinite time constant [52]. Infinite time constants at each node do not generally reflect the dynamics of biological systems. For example, proteins degrade at different rates in transcriptional regulatory networks. With finite-dimensional linear dynamics all networks, except a set of parameters of zero measure, are controllable with a single input. Because the model above omits the intrinsic nodal dynamics that arise due to processes that have nothing to do with network topology, it requires self-links to be added where appropriate. However, in adding self-links to each node to represent intrinsic dynamics, all nodes in the network become matched nodes [52]. This would imply that any network can be controlled with a single input effecting the power dominating set of nodes (nodes which do not have an input from another node). This raises the issue of how appropriate it is to apply the concept of structural controllability to real complex networks.

***Pareto front optimality***

The Pareto front concept from economics and engineering is used to find designs that are the best trade-offs between different requirements and has been used to explore bacterial gene expression [22]. If you consider two phenotypes, A and B, if A is better at all tasks than B then B will be eliminated by natural selection. Repeating this process for all possible phenotypes leaves the Pareto front which is the set of phenotypes which cannot be improved at all tasks at once. The Pareto front can be calculated as the line that connects two archetypes in which an archetype is defined as a phenotype which is best at a given task. A phenotype that has more than two tasks will be characterized by a higher dimensional geometry, with a line connecting the archetype for each task. For example, 3 tasks produces a triangle and four tasks a tetrahedron. The activity of 1600 promoters in *E. coli* during growth indicates two distinct clusters of genes with one cluster consisting of mostly growth genes and the other primarily stress and survival genes. Recording the percentage of total promoter activity over time showed that expression patterns existed in a one dimensional line, even when recorded over 4 environments. Over time, gene expression gradually moves along the line from cluster 1 to 2 [53].

***Principle Component Analysis***

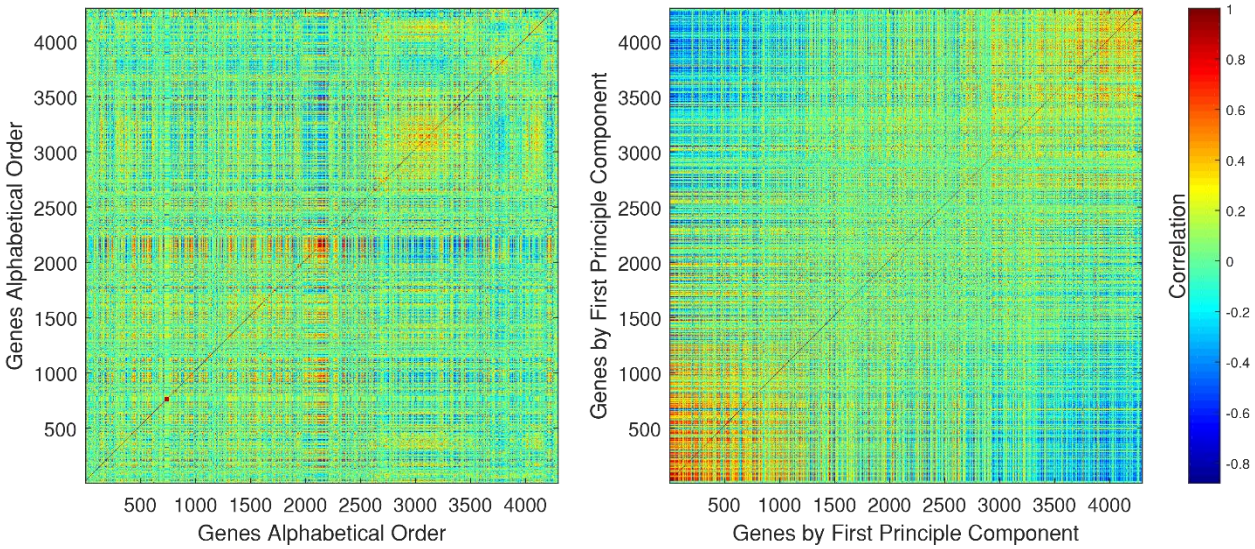



*Figure 4: The correlation matrix of microarray data with expression of 4320 genes from 466 experiments. Highly Correlated genes appear in red and anti-correlated genes appear in blue. On the left, genes are arranged alphabetically. On the right, genes are sorted by their contribution to the first principle component after single value decomposition, revealing two clusters of anti-correlated genes [42].*

The finding of distinct clusters of gene expression of stationary or exponential growth genes is consistent with singular value decomposition of microarray data (Figure 4) [42]. This reorders genes and experimental conditions according to their main axes of variation. The result is two globally anti-correlated gene clusters in which one cluster is preferentially expressed during exponential growth and the other during stationary phase [42]. There is a strong association between sigma 70 and the global pattern of anti-correlation. This is not surprising given that most housekeeping genes that are transcribed in exponential phase are under sigma 70 promoters. However retaining only operons known to be transcribed with sigma 70 is not enough to suppress the anti-correlations. The majority of correlated pairs of genes do not share a common transcription or sigma factor and therefore regulation of operons by the transcription factor network are not enough to explain the patterns of gene expression observed in micro array data.

***Linear Regression for Estimating Epistasis***

Epistasis occurs when the whole system is not equal to the sum of its parts. Specifically, if there are two mutations A and B, the expected fitness of the strain with both mutations AB should be the sum of the fitness effects from A and B [54].

$$\mathbf{f_{AB}} = \Delta\mathbf{f_A} + \Delta\mathbf{f_B} + \mathbf{f_{wt}}$$

When this is not the case, it is referred to as epistasis. If the fitness is higher than expected, it is synergistic epistasis, if it is lower than expected it is antagonistic epistasis. This implies that the fitness effects of single mutations are context dependant. That is that the mutation **A** has a different change in fitness (**Δf**) in the wild-type (**wt**) context than in a context with mutation **B**. If the sign of the effect of **A** changes in these two contexts, (for example, in **wt**, **A** causes an increase in fitness but in **B**, **A** causes a decrease in fitness) this is referred to as sign epistasis.

With a complete dataset of phenotypes for all combination of genotypes (of size $2^n$ where n is the number of genes being perturbed), linear regression can be used to calculate epistasis coefficients (β) for all genotypes [55]. In matrix form, this is represented by the following equation:

$$\bar{y} = \mathbf{X}\bar{\beta} + \bar{\epsilon}$$

Here y are the $2^n$ measured phenotypes, ε a residual noise term and **X** is a $2^n * 2^n$ matrix that follows a recursive form [55]:

$$\boldsymbol{X}_{n+1} = \begin{pmatrix} \boldsymbol{X}_n & 0 \\ \boldsymbol{X}_n & \boldsymbol{X}_n \end{pmatrix} \text{ where } \boldsymbol{X}_0 = 1$$

As long as we have all $2^n$ measurements for y, we can solve for β [56]:

$$\bar{\beta} = \mathbf{X}^{-1}\bar{y}$$

We can also approximate data with fewer epistasis coefficients. This is done by reducing **X** to eliminate columns that refer to epistatic orders we wish to exclude. For example if we wish to only consider pairwise epistasis r = 2, yet if we wish to consider all epistatic interactions r = n. To reduce **X** we use the following equation [55]:

$$\widehat{\mathbf{X}} = \mathbf{XQ}$$

Here **Q** is an $2^n$*m identity matrix where m is the number of epistatic terms up to r and is given by [55]:

$$\mathrm{m} = \sum_{i=0}^{\mathrm{r}} \binom{\mathrm{n}}{i}$$

The linear regression is then solved by [55]:

$$\hat{\beta} = \left(\widehat{\mathbf{X}}^T\widehat{\mathbf{X}}\right)^{-1}\widehat{\mathbf{X}}^T\bar{y}$$

We can then use the fewer number of coefficients to recalculate the expected fitness measurements and compare them to the observed fitness measurements. This allows us to estimate how much the higher order epistasis influences the fitness of the perturbations. Importantly however, it does not account for non-linearities in the data. No epistasis can also be defined as multiplicative rather than additive (where it can be useful to normalize fitness of the reference strain to 1) [54].

$$\mathbf{f_{AB}} = \mathbf{f_A} * \mathbf{f_B} * \mathbf{f_{wt}}$$

When linear regression is performed on non-linear data with high order epistasis two effects will occur on the calculated fitness vs observed fitness plot. First, the trend will bend due to the non-linear effects. Second the trend will spread or become noisy due to the higher order epistatic effects [56].

élément sous droit, diffusion non autorisée

*Figure 5: Patterns of nonadditivity for increasing epistasis and nonlinear scale. Additive-Predicted phenotypes (x-axis) vs observed phenotypes (y-axis). Nonlinearity increases from left to right plots, with the leftmost plots being a completely linear scale and the rightmost plots being a highly non-linear scale. Epistasis increases from bottom to top rows with the bottom row of plots representing no epistasis, and the top row of plots representing high epistasis [57].*

***Logical Modelling***

Logical modeling represents regulatory interactions as logic gates and has discrete states for each gene. The simplification to discrete instead of continuous states makes it easier to analyse large biological networks [58]. These discrete states can be either Boolean or multi-level [59]. State transitions occur at each time step [46], which updates the logical network by using the current states to calculate the new states of the downstream nodes. The dynamic behavior of the logical network is represented in a state transition graph, [59]. Stable states and oscillatory behaviours can then be extracted from the state transition graph, as nodes with no outgoing arcs and strongly-connected components respectively.

élément sous droit, diffusion non autorisée

*Figure 6: Logical model implemented in GINsim. a) Three nodes representing protein A, B, and the complex AB. A is always on unless inhibited by AB. AB is on only if both A and B are on. B can exist in 3 levels, it is at 0 if repressed by AB, 2 if activated by A, and 1 if neither of those is true. b) the combination of all trajectories possible by the logic rules creates the state-transition graph. Regardless of the starting state, the system will be attracted to one of the highlighted blue states in a strongly connected component, where it will begin to oscillate [46].*

Logical models are versatile because nodes can represent almost anything [46]. However they rely on existing knowledge to construct. This assumes that there is a known structure for a regulatory network and that the transcriptional regulatory network depiction is accurate. This can be overcome with various network inference techniques, which use experimental evidence to determine the regulatory network structure [60]. In the case of *E. coli*, regulatory network representations currently exist [45] [44], although these are continuously being updated with new data and improving techniques [61].

## 1.4 *Large scale control and programming of gene expression using CRISPR*

***Matthew Deyell[1], Sandeep Ameta[1], Philippe Nghe[1]****

*Affiliations*

*[1]Laboratoire de Biochimie, CNRS UMR8231, Chimie Biologie Innovation, PSL Research University, ESPCI Paris, 10 Rue Vauquelin, 75005, Paris, France*

*Correspondence to: philippe.nghe@espci.fr*

**ABSTRACT**

The ability to control and regulate gene-expression in biological as well as synthetic systems has allow the better understanding of *gene* to *function* relationship. Several systems have been exploited to achieve this such as Zinc fingers, TALE (Transcription activator-like effectors), siRNAs (small-interfering RNAs). However, recent advances in Clustered Regularly Interspaced Short Palindromic Repeats and Cas9 (CRISPR-Cas9) have overshadow them due to its high specificity, compatibility with many different organisms, and flexibility in targeting multiple gene at once. These features make CRISPR-Cas9 an efficient system for large scale gene perturbation studies. In this review we summarize state-of-the art for CRISPR-Cas9 technology and their use in gene knock-out, knock-down/up screenings. We feature the recent studies where CRISPR-Cas9 ability to create large scale perturbation libraries has been combined with single-cell sequencing, which show their adaptability and efficiency in addressing biological problems at a single-cell level. Additionally, we also highlights the application of CRISPR-Cas9 system in building synthetic logic circuits to engineer and program cell for achieving better control and regulation.

**Contents**



Increasing, decreasing or abolishing the expression of individual genes and observing the impact of such perturbations on various cellular processes is the basis of powerful inference methods for gene function. Designing synthetic cells also requires to master the control of gene expression. Up to now, the major approach to modifying gene expression has been through homologous recombination, in which a selection marker (such as a gene for drug resistance) is flanked by targeted sequence from the genome of interest. This construct is then transformed into the cell, and a selective drug selects only cells have integrated the desired sequence into their genome. While effective, the site of homologous recombination is generally not well controlled, the methodology is time consuming, and extremely challenging for cells which do not possess efficient homologous recombination pathways. This limited most perturbation studies to a few model organisms such as *E. coli* and *S. cerevisiae*.

Our ability to perform gene perturbation studies in a large diversity of organisms beyond the traditional models has been greatly enhanced with the discovery and development of proteins able to bind to DNA at specific sequence. Initially, these consisted of Zinc fingers [62] and Transcription activator-like effectors (TALE) [63]. These DNA binding domains fused to nuclease domains cause double strand DNA breaks at specific sites within the genome. The double strand breaks can then be repaired by the non-homologous end joining (NHEJ) pathway, leading to small insertions or deletions. When these occur within a gene, a frame shift can occur, making the gene inoperative. This can easily generate homozygous knock-out mutants, even in diploids without performing selective breeding, as identical targets are disrupted on

each allele within each cell. Additionally, other effector molecules such as the tetrameric Herpes Simplex Viral Protein 16 transcription activator (VP64) domain [64] or the Krüppel associated box (KRAB) domain can be fused to Zinc Fingers or TALE, instead of nuclease domains, to respectively knock up or knock down expression of specific genes rather than knock them out completely. These fusion proteins provide a greater flexibility in both the direction (knock up or down) as well as the intensity of the perturbation, and function as synthetic transcription factors [65] [66] [67] [68] [69].

The discovery of a programmable endonuclease in the form of the CRISPR-Cas9 system [70] has overshadowed both Zinc fingers and TALEs because it possesses the same functionalities while facilitating the systematic perturbation of entire genomes [71] [72] [73] [74] [75]. This is because the DNA binding sequence specificity of CRISPR-Cas9 is determined by a short RNA rather than the protein sequence of Zinc fingers and TALEs, which has to be engineered and screened for each new target. Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) were originally identified as an adaptive bacterial defense system against phage. This system allows a single protein (Cas9) to target many different phage genomes simultaneously while still remaining highly specific. It does this through the help of a short targeting CRISPR RNA (crRNA) and an accessory trans-acting RNA (tracrRNA) which recruits the endonuclease Cas9 to the specific sequence (Figure 1A). The crRNA binds to the host DNA as well as to the tracrRNA, which associates with the Cas9 protein. In the most commonly used CRISPR-Cas9 system, from *Streptococcus pyogenes*, the crRNAs come in an array of 30 bp targeting sequences flanked by 36 bp semi-palindromic repeats. The crRNA array must be cut into individual units by RNase III to become functional [76]. The crRNA-tracrRNA-Cas9 complex is able to bind DNA only when the target sequence is located next to a so-called Protospacer Adjacent Motif (PAM) sequence [77]. This complex then induces double stranded breaks within the target sequence 3 nucleotides from the PAM site [78]. For *S. pyogenes* Cas9 the PAM sequence motif is 5'-NGG-3', where N stands for any nucleotide. This PAM sequence is crucial as it provides self-immunity from CRISPR-Cas9, and notably prevents Cas9 from destroying its own crRNA array.

The complete process of targeting genes using CRISPR-Cas9 has been streamlined by the engineering of a small guide RNA (sgRNA) which both mimics the crRNA-tracrRNA complex [79] [80] and removes the need for RNase III (Figure 1A). The sgRNA consists of a 20 nt targeting region, a 42 nt Cas9-binding hairpin structure, and a 40 nt transcription terminator from *S. pyogenes* [77]. Similar to the crRNA-tracrRNA complex, programming of the target by the sgRNA depends on complementarity to a 20 nt region adjacent to a PAM site. While the crRNA and sgRNA have a targeting sequence length of 30bp and 20 bp respectively, in both cases the specificity is mostly determined by the PAM sequence and the 12 'seed' nucleotides immediately following it. Since specificity is determined by this limited number of nucleotides, off-target effects may occur in large genomes [77]. However, potential off-target sites which differ from the target site by up to 3 bp generally show only minimal cleavage unless they have perfect complementarity to the terminal 8 bp within the 'seed' region [71].

**2. CRISPR-Cas9 Gene knock-out screening**

A functional CRISPR-Cas9 system, comprising the Cas9 protein, a sgRNA targeting region, and a PAM sequence, results in double strand breaks by the endonuclease activity of Cas9 (Figure 1B). If host cells contain a NHEJ system, these breaks will be repaired resulting in deletions smaller than 20 bp [71]. These deletions lead to frame shifts and loss of function of the targeted protein. However, not all cells will display complete loss of function, despite expressing both sgRNA and Cas9. Typically one third of the cells

continue to express the targeted gene [81], which is in line with predictions that one third of repairs made by the NHEJ system should produce in-frame mutations [81]. While these in-frame mutations may still be able to create a full-length protein, they are not necessarily neutral but may also cause partial loss-of-function or gain-of-function, impacting the overall fitness of the protein [82] [83] [84].

The capability of CRISPR-Cas9 to induce knock-out phenotypes in a programmable fashion makes it uniquely suited to create genetic knock-out libraries which can be screened using positive or negative selection. For mammalian cells, this is accomplished by creating a pool of lentiviruses, each of which expresses a sgRNA targeting a specific gene. This has allowed genome-wide knock-out studies in human cell lines [75] [71]. The sgRNA is advantageously exploited to act as a barcode; counting the number of each barcode with high-throughput sequencing to know the proportion of each sgRNA in the population. The change in the frequency of sgRNAs after selection results in a 'fitness score' for the gene targeted by the sgRNA [71]. Frequencies can also be measured over time to fit growth curves for each knockout-mutant. Furthermore, targeting the same gene with multiple sgRNAs can increase the robustness of the fitness score estimate [85]. Measurements of fitness are not only limited to growth selections. Knockout libraries have been stimulated with lipopolysaccharide (LPS) and selected by flow cytometry for cells that fail to induce the inflammatory cytokine Tnf [72]. This approach recovered known key regulators, validated new regulators, and identified novel distinct regulatory modules [72]. Multiplexed knockout libraries are also easily created by delivering multiple sgRNAs to each cell [85].

Beyond single gene perturbations, genetic interaction scores can be estimated by the difference in fitness observed between the single and double sgRNA knock-out cells. Array-based oligonucleotide synthesis has been used to create dual sgRNA libraries covering $10^5$ gene pairs [85]. Interestingly, libraries of sgRNA containing single genes and all gene-gene combinations can be made at once, by introducing sgRNAs that do not target any location within the genome. Indeed, when a non-targeting RNA is paired with a targeting sgRNA during pooled cloning, this pair is equivalent to a single perturbation. Using this approach, a total of 152 synthetic-lethal (negative) genetic interactions, and 10 positive genetic interactions were identified in HeLa, A549, or 293T cells [17].

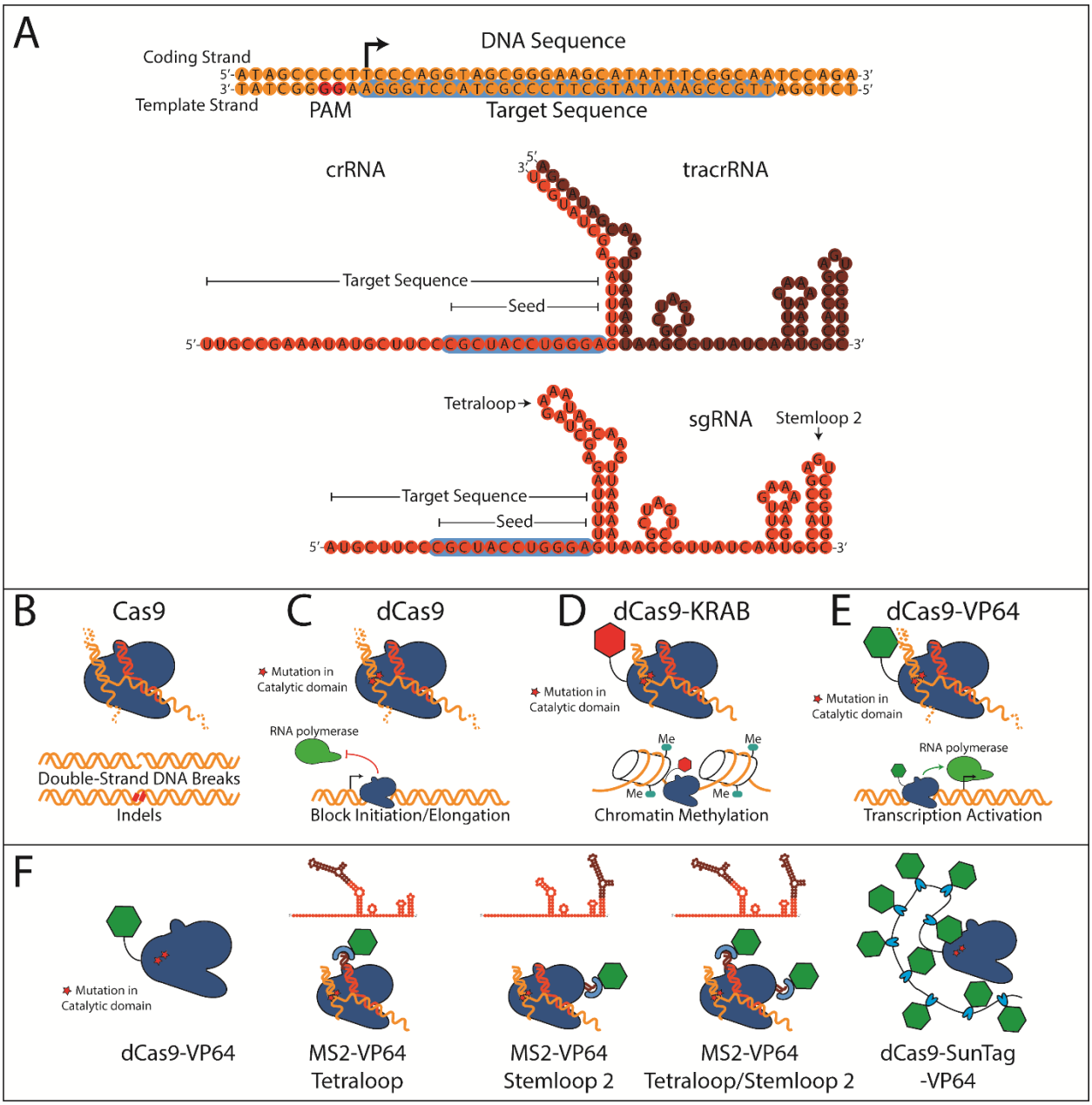



*Figure 7: CRISPR-Cas9 Variants for Controlling Gene Expression. A) When searching for targets for CRISPR-Cas9 applications, PAM sequences (5'-NGG-3') must be identified in the region of interest (yellow). These PAM sequences can be located in either the coding strand or the template strand. However, for dCas9 applications blocking transcription elongation, it is recommended to find PAM sequences in the template strand for enhanced repression. The 30 nucleotides preceding the PAM sequence become the spacer sequence for the crRNA or sgRNA. When using sgRNA, only the first 20 nucleotides preceding the PAM sequence are typically used. B) The native Cas9 causes double strand breaks in DNA. This can be useful for knock-out screening as they will typically be repaired as small insertion/deletions (Indels) which cause frameshift mutations and abolish gene expression. C) The catalytic sites of Cas9 can be inactivated resulting in a dead version of the enzyme which retains its DNA binding capacity. This in effect transforms Cas9 into an RNA programmable DNA binding domain. By targeting promoter or protein coding sequence of a gene, transcription initiation or elongation can be blocked respectively. D) The dCas9 protein can be fused to transcriptional activator domains such as VP64 to activate transcription. E) Similarly the chromatin state of a gene can also be changed by localizing effector domains; dCas9 protein can also be fused to KRAB effector domains to methylate histones and block transcription. F) Effector domains can be localized to DNA sequence of interest through multiple ways. In addition to fusing them directly to the dCas9 protein, they can be fused to MS2 RNA binding domains and localized to the sgRNA instead. MS2 recognition*

*sequences can be added to multiple locations on the sgRNA including on the tetraloop and the stem loop 2 sites. Multiple MS2 recognition sequences can be added to a single sgRNA to increase the number of effector domains recruited by the CRISPR-Cas9 system. Finally, the SunTag approach links a chain of multiple repeating peptide recognition sites to the end of dCas9. An effector domain fused to a small antibody sequence recognizes and binds to the peptide sequence in the chain, allowing for recruitment of upwards of 10 effector domains to a single dCas9. The increased number of effector domains localized to the sequence enhances their effect on transcription.*

### 3. CRISPR-dCas9 Gene Knock Down / Up screening

Genetic knock-out screening has multiple limitations. Firstly, the NHEJ system often produces short, in-frame indels resulting in approximately one-third of the cells that continue to express the gene of interest [81]. Secondly, the irreversible nature of frameshift disruptions limits their use in targeting essential genes [73]. And thirdly, double strand breaks caused by CRISPR-Cas9 invoke the SOS response [86], causing cytotoxicity [73] [87]. To overcome these issues, Cas9 can be rendered catalytically inactive, effectively transforming it into an RNA programmable DNA-binding protein. This allows for transcriptional regulation of a target sequence by blocking the RNA polymerase, without modifying the genetic sequence [80] [88]. The catalytically dead Cas9, termed dCas9, contains two point mutations in the RuvC1 nuclease (D10A) and the HNH nuclease domain (H840A) [80] [77] (Figure 1C). This method of gene regulation is often referred to as CRISPR interference (CRISPRi) and has numerous advantages over other methods of transcriptional regulation. It is functional in many organisms including both bacteria and mammalian cells, it requires only a single protein species (dCas9), and can target any gene of interest by customizing the small targeting RNA (sgRNA) [77]. Compared to earlier transcriptional regulation methods, such as Zinc fingers and TALE, only the 20 nt gene complementary region of the sgRNA must be changed rather than an whole additional enzyme [77]. This allows for cost-effective large scale perturbation experiments to be performed. These large scale screens are possible with RNAi as well. However, CRISPRi has been shown to have little off target effects compared siRNA [77], given that the latter may repress hundreds of transcripts due to imperfect matching to mRNA 3'-UTRs (untranslated regions), which may occur redundantly across many genes [82] [89] [90] [91] [73].

Transcriptional silencing with CRISPRi allows up to 1,000-fold repression [80]. Repression strength can be tuned by adjusting the targeting location of the sgRNA along the gene. Within a gene, 10 to 300 fold repression can be achieved by designing the sgRNA to be complimentary to the coding strand of DNA, while a sgRNA complimentary to the template strand results in little to no effect [80]. In contrast, targeting the promoter region of a gene will significantly perturb gene expression regardless of which strand is targeted, with a maximum of ~100 fold repression at the -35 region of the promoter [80]. Combining two sgRNAs which target the same gene shows multiplicative effects as long as their targets do not overlap [80]. However, these repression efficiencies are sensitive to mismatches in sgRNAs [77] as even a single mismatch at the 3' end of the sgRNA's targeting sequence will substantially decrease the CRISPRi activity [73].

Several options are available to temporally control dCas9 activity. Transcriptional repression can be made inducible and completely reversible by placing dCas9 under the control of an inducible promoter to trigger loss-of-function on demand [80] [77] [81]. Repression was observed within 10 minutes of adding an inducer and could be reversed within 50 minutes by washing, likely corresponding the time necessary for dilution of dCas9 and the sgRNA during cell growth and division [80]. This technique has allowed temporal regulation of transcription factors in induced pluripotent stem cells (iPSC) [81]. The leaky expression from inducible promoters can be further controlled by splitting dCas9 and fusing it to FRP-FKBP

dimerization domains, which associate upon rapamycin addition [92]. There was no significantly detectable dCas9 activity in the absence of rapamycin [93]. Furthermore, the rapamycin induced complex was found to have reduced off-target effects compared to a non-split dCas9 protein. Alternatively dCas9 can be conditionally stabilized by fusion to a DHFR (dihydrofolate reductase) domain, which directs proteasomal degradation of the Cas9 fusion protein in the absence of TMP (trimethoprim) [94]. Further control can be acquired by combining CRISPR and antisense RNA (asRNA) systems [95]. The de-repression of a targeted gene can be achieved by expressing an asRNA which sequesters the sgRNA allowing transcription resuming independently of dilution dynamics. The amplitude of de-repression was found to be proportional to the strength of RNA-duplex formed between asRNA and sgRNA [95].

In addition to directly blocking transcription initiation and elongation, dCas9 can be used to localize effector domains that either repress or activate transcription [96] [73]. Transcription in mammalian cells can be repressed by fusion of dCas9 to KRAB domain [96] (Figure 1D). This has been used to downregulate 107 individual genes or pair of genes of chromatin-regulation factors in human cells, revealing a functional map of chromatin regulation [88]. The relative number of sgRNAs between the induced and uninduced samples are quantified by next-generation sequencing (NGS), and used as a proxy for relative cell fitness [88]. While more than 75% of the sgRNAs were able to repress the gene expression of targeted genes, 30% of double knock-downs were able to repress the targeted genes. Additionally the repression efficiency was overall lower than RNAi with a median expression of approximately 50% compared to 30% with RNAi [88] [97]. However unlike RNAi, this approach can be exploited for gain of function screening with CRISPRa. CRISPR activation (CRISPRa) uses a dCas9 fused with a C-terminal tetrameric VP64 domain [64] (Figure 1E). Transcriptional induction peaks when targeted between 50-400bp upstream of the transcriptional start site [73].

An alternative to fusing VP64 directly to dCas9 is to engineer the sgRNA as an RNA scaffold for VP64 recruitment. Indeed, a fusion between VP64 and the MS2 bacteriophage coat protein can be recruited to hairpins located on the end on the sgRNA [74] (Figure 1F). Recruiting to the tetraloop or the stem loop 2 of the sgRNA (Figure 1 A, F) results in a 3 to 5-fold higher expression level than dCas9-VP64 respectively [74]. Expression can be enhanced up to 15 fold by recruiting MS-VP64 to both positions on the sgRNA in combination with a dCas9-VP64 fusion protein [74]. As the activation strength seems to increase with the number of effector domains that can be recruited, atechnique named SunTag [98] has been developed to provide up to 10 effector domains (Figure 1 F, rightmost panel) with a 10 to 50 fold increase in gene activation compared to a simple dCas9-VP64 fusion [98]. Finally, temporal control of activation can be obtained by expressing effector domains as peptides distinct from the dCas9, and only dimerize upon addition of a chemical ligand [99]. A note of caution is that the binding of Cas9 may affect nucleosome positioning and binding of nearby transcription factors [100].

### 4. Identification of Regulatory Regions with CRISPR-Cas9

Unlike RNAi techniques, CRISPR-Cas9 is not limited to targeting only transcribed regions of the genome. The CRISPR-Cas9 system can alter non coding genomic sequences at a high-throughput [101] to identify essential regulatory elements [100]. To map the genomic elements required for expression of a single gene, the latter is first replaced with a GFP marker. Here, sgRNAs are designed to target multiple regions across the gene locus, including non-coding cis-regions. Cells with low GFP levels are sorted and sequenced, such that Cas9 disruption reveals critical regulatory regions [102]. This allowed to discover novel promoter and enhancer regions for Nanog, Rpp25, Tdgf1, and Zfp42 [102]. Interestingly, when this

approach was applied to POU5F1, nearly 40% of cis-regulatory sequences corresponded to promoters of other genes, suggesting long-range chromatin interactions [103].

Alternatives to fluorescence sorting are proliferation based assays. GATA1 and MYC transcription factors were investigated for regulatory regions by creating a total of 98,000 sgRNA for all possible locations around the two genes [104]. Cells expressing KRAB-dCas9 were then infected with sgRNA library and the quantity of each sgRNA was sequenced before and after the growth. The region with the strongest depletion of sgRNA corresponded to the transcriptional start site. Additionally, three distal regulatory elements were identified for GATA1. In MYC, 7 distal regulatory elements were located from sgRNAs depletion, while 2 other sites corresponded to sgRNA enrichment [104]. Similar screening of non-coding regions can be done by knock-down or knock-up perturbations instead of knock-out. This was done by replacing Cas9 with KRAB-dCas9 and P300-dCas9 respectively (P300 being an alternative to the VP-64 domain) [105]. This enables to test loss and gain of function with the same library of sgRNAs. This may be useful because active regulatory elements may not be identified in gain of function screens while repressed elements may not be identified in loss of function screens [105].

Finally, CRISPR-Cas9 can be used to specifically modify the sequence of regulatory region. Homology directed repair (HDR) with the CRISPR-Cas9 system can create a library of precisely determined genotypes [100]. Because of the ease of multiplexing and the predictability of HDR genome editing with CRISPR-Cas9, it is ideally suited for high-throughput screening of single nucleotide polymorphisms (SNP). High-throughput single-nucleotide polymorphisms sequencing (SNPs-seq) utilizes CRISPR-Cas9 to characterize allele-dependent transcriptional regulation in prostate cancer risk loci [106]. The sgRNA can be designed to target specific SNP regions to evaluate their effect on gene expression. This identified allele specific effects, in which alleles containing a SNP at a risk loci had different impacts on gene expression with one allele no longer acting as a cis-regulatory element.

**5. Transcriptional profile screening with Perturb-Seq**

The development of droplet-based single-cell RNA sequencing (scRNAseq) techniques [17] [18] [107] allows for the transcriptional profiling of tens to potentially hundreds of thousands of individual cells. This technology has been paired with the CRISPRi system to generate large libraries of genetically perturbed cells [108] [20] and analyse the effect of those perturbations on a large set of genes in a highly parallel manner (Figure 2). In the scRNAseq protocol, single cells are encapsulated into nano-litre droplets containing reagents for reverse transcription and a uniquely barcoded hydrogel bead carrying primers. The reverse transcription in droplets generates droplet-specific barcoded cDNAs which are appended with sequencing adaptors before analysis by high-throughput sequencing platforms. Cell-specific sequencing reads are then de-convoluted during bioinformatics analysis of the bar-codes. As the sgRNA for CRISRPi are also associated with a specific bar-code at the level of the lentivirus vector, scRNAseq associates single cell transcriptomes with CRISPR perturbations through the droplet bar-code (Figure 2E). This approach is currently limited by the sequencing depth rather than the potential size of the CRISPR-Cas9 knock-down library or the microfluidic operations. Finally, multiple genes can be perturbed per single cell by controlling the multiplicity of infection of the lentivirus carrying the sgRNA [20]. This allows for the identification of non-linear effects (epistasis) caused by interactions between multiple genes on transcriptional profiles.

Perturb-Seq has been used to study the inflammatory response of bone marrow dendritic cells in response to lipopolysaccharide (LPS) [20]. In this response, approximately 2000 genes are induced through dozens

of transcription factors in an asynchronous response consisting of at least two sub-types, a heterogeneity which can only be revealed by a single cell approach. Dixit *et al*. analyzed 24 transcription factors, grouping them into modules with similar regulatory effects. They also grouped genes by their response to perturbations into genetic programs, and found that transcription factor modules regulated specific gene programs consistent with their known functions [20]. The targets identified for their transcription factors was also supported by chromatin immunoprecipitation sequencing (ChIP-seq) from bulk populations, and correctly predicted the targets and logic of transcriptional repressors [20]. This demonstrated the ability for Perturb-seq to recover at a large scale the genes, processes, and states regulated by transcription factors.

Perturb-Seq has also been used to analyze the mammalian response to unfolded protein [108]. Genes that contribute to ER homeostasis were first identified using two genome-wide CRISPRi screens. These screens were used to determine a subset of genes to investigate at the single cell level in order to define their functional relationships [108]. A single cell approach is crucial here due to subpopulations of cells within the cell types identified in bulk approaches [108]. The high-throughput nature of droplet-based single cell RNA sequencing also helps to correlate gene expression profiles to infer transcriptional programs [108]. Adamson *et al*. found that three endoplasmic reticulum transmembrane sensor proteins (PERK, ATF6, and IRE1) had both distinct and overlapping gene expression programs in the unfolded protein response. Additionally, the ability to perform single cell RNA sequencing leads to the identification of distinct subpopulations for a same perturbation of the HSPA5 gene. This observation that seemingly identical cells would have differential responses would be lost in a bulk RNAseq screen.

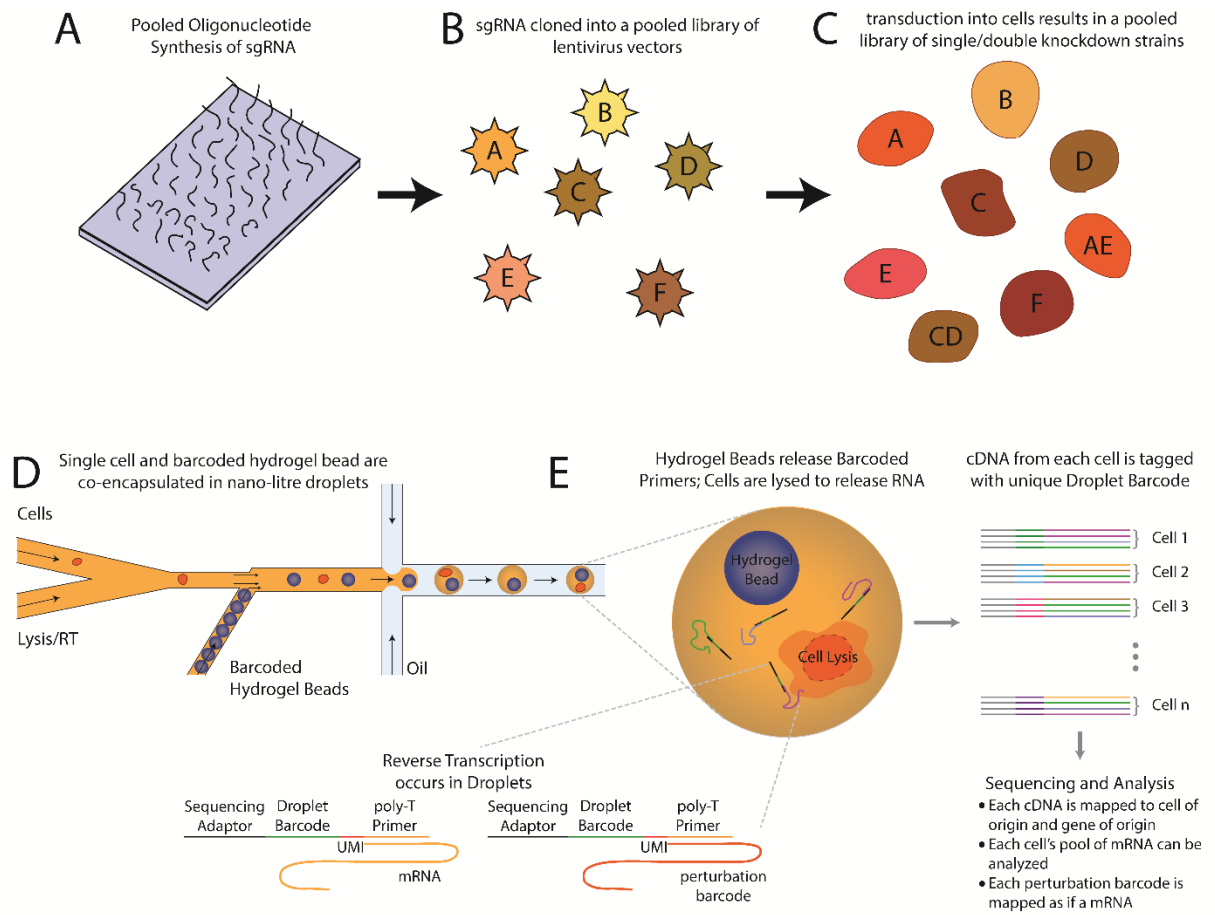



*Figure 8: Massively-Parallel screening of CRISPR-Cas9 perturbation libraries with Perturb-Seq. A) Oligonucleotides containing thousands of target sequences for CRISPR-Cas9 can be synthesized in pools on microarray chips. B) These pooled oligonucleotides can then be used to create sgRNA sequences contained within lentivirus vectors for CRISPR-Cas9 perturbations. C) The library of CRISPR-Cas9 vectors can then be transduced into a cell line. By controlling the multiplicity of infection, it is possible to have single, double, and even triple knock-down strains. D) Cells that have been perturbed by CRISPR-Cas9 are encapsulated into nano-litre volume droplets, with each droplet containing at most one cell and a hydrogel bead bearing reverse transcription primers with a DNA barcode unique to that hydrogel bead. E) Droplets act as reaction chambers to perform reverse transcription. All of the RNA from a single cell is converted into cDNA bearing the barcode from the same hydrogel bead, as well as a unique molecular identifier (UMI) ( [109]). When the cDNA from all cells is sequenced, the barcode allows all the cDNA from a single cell to be associated with a single droplet, and the UMI corrects for any application bias and allows quantification of the original RNA for each gene.*

Careful note must be taken when processing the sgRNA lentiviruses used in Perturb-Seq. Recombination between vectors can scramble the association between sgRNA and perturbation barcodes [108]. Two viral genomes are packaged in each lentiviral vector and the reverse transcriptase can switch between the two templates during provirus synthesis [110]. The larger the sequence length between the sgRNA and its associated barcode, the greater the chance for this recombination to shuffle the sgRNA and barcodes. Xie *et al*. found that recombination accounted for up to 50% of all reads [110]. While this did not lead to false positive hits, it did reduce the signal to noise ratio. This can compound with the already noisy single cell RNA sequencing data. An alternative to having a separate perturbation barcode is to sequence the sgRNA directly as barcodes. This approach is taken in so-called CROP-seq [111]. Here, the guide RNA becomes part of the puromycin-resistance mRNA and is detectable by RNA-seq protocols which use poly-A enrichment. Functional sgRNA are still produced from a hU6 promoter. This approach was validated

against T-cell-receptor induction in Jurkat cells and validated against bulk RNA-seq and flow cytometry, providing a solution to recombination between sgRNA and separate perturbation barcodes [111].

While the aforementioned techniques utilize a droplet-based approach to perform single cell RNA sequencing, CRISPR perturbation libraries are compatible with other single-cell approaches as well. In CRISPR-Seq, the vector carrying the sgRNA also contains a fluorescent selection marker in addition to a sgRNA associated DNA barcode [19]. This allows single cells to be sorted by fluorescence activated cell sorting (FACS) into 384 well plates [112]. This approach identified the development regulatory genes Cebpb and Irf8 as controlling cell differentiation commitment in myeloid lineage. When Cebpb was perturbed, cells favored dendritic cell lineages over monocytes and expressed high levels of Irf8 [19]. Given the high plasticity of myeloid cells, such an observation would be difficult without a single cell approach.

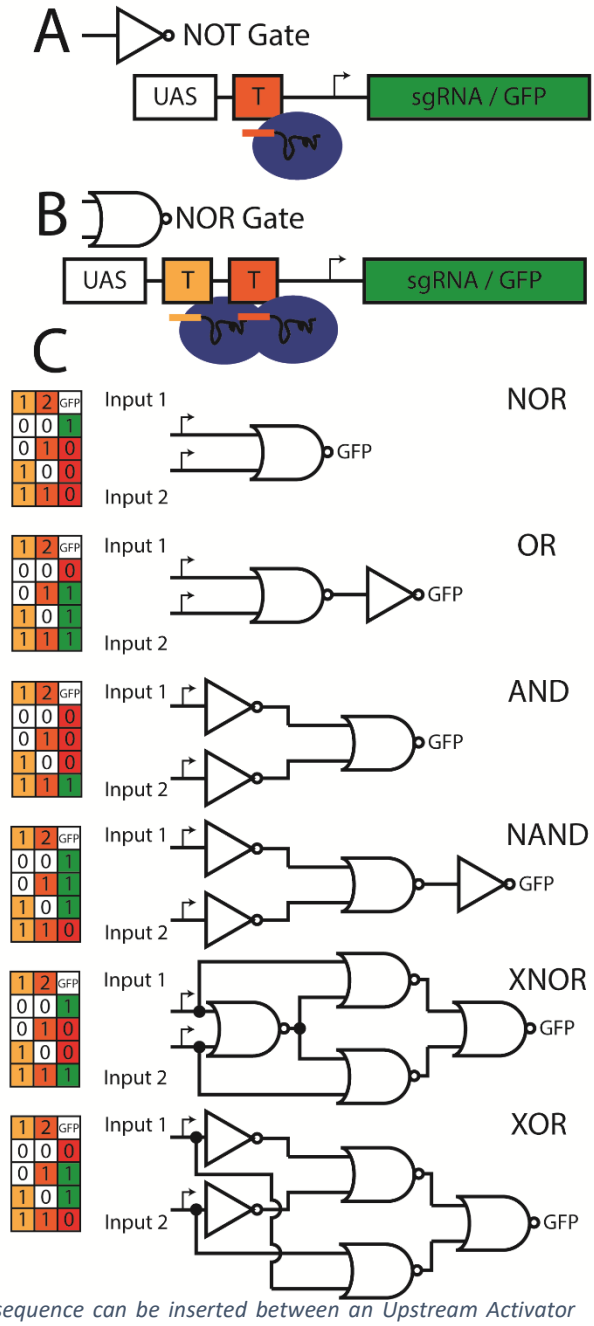
## 6. Synthetic logic circuits with CRISPR-Cas

Well characterized, orthogonal synthetic transcription factors would allow engineering of novel regulatory networks to control living cells [113]. Creating NOR or NAND logic gates from orthogonal synthetic transcription factors are in principle sufficient to build all possible logic gates [114]. Indeed, combinations of these logic gates can be used to express all possible truth tables in Boolean logic [115]. Transcription factors have been used to create NOR gates by placing two promoters in series to drive a transcriptional repressor [116] [117]. Zinc fingers and TALE have also been used to make synthetic transcription factors, however engineering these proteins can be difficult and expensive as each protein must be individually design and tested [118]. CRISPR-dCas9 provides an orthogonal DNA binding protein which can be programmed to target multiple sequences by changing only a short sequence of targeting RNA.

The programmable nature of CRISPR-dCas9 has allowed it to be used to scale-up the regulatory network, and build multiple component circuits by inhibiting initiation or elongation by RNA polymerase (Figure 3). Because multiple guide RNAs can direct dCas9 to target multiple sites within a single cell, logic gates can be made by having dCas9 target different guide RNA (Figure 3). A single sgRNA targeting the promoter of a gene acts directly as a NOT gate (Figure 3 A). This is because the sgRNA must not be expressed for expression of the gene of interest to occur. A NOR gate can be constructed by having two unique sgRNAs target the promoter of a gene, so that either sgRNA will shut transcription off (Figure 3 B). This can be converted into an OR gate by making the two sgRNA repress the expression of a third sgRNA, which itself represses the gene of interest. By using up to five NOR gates and seven sgRNAs, NOR, OR, AND, NAND, XNOR and XOR gates have all been created (Figure 3 C) [113].

CRISPR AND logic gates can also be used to drive gene expression only in specific cell lines. For example, by placing Cas9 under a bladder-specific promoter hUP II, and a sgRNA under human telomerase reverse transcriptase promoter hTERT, specific effectors will only be expressed in bladder cancer cells [119]. Other specific effectors used included hRluc to detect cancer cells, hBAX to induce apoptosis, p21 to arrest growth, and E-cadherin to reduce cell mobility. Each of these effectors was under the control of a CMV promoter with a LacO site to repress transcription. These effectors were coupled to the CRISPR-Cas9 system by having the sgRNA under hTERT target LacI expression, a lactose responsive transcriptional

repressor from *E. coli* that binds to LacO. When both Cas9 and the sgRNA are expressed, lacI is repressed allowing for transcription of the effector to occur [119].




*sequence can be inserted between an Upstream Activator Sequence (UAS) and a promoter that corresponds to a specific sgRNA. A) A NOT logic gate targeted by a single sgRNA will prevent expression. B) Adding two unique targeting sequences creates a NOR gate, in which neither of the sgRNA can be expressed for the gene to functional. C) By linking together NOT and NOR gates (having them express further sgRNA), other logical functions can be created.*

## 7. Integrated Genetic circuits with CRISPR-Cas

Synthetic CRISPR genetic circuits can be constructed by combinations of NOT and NOR gates to generate distinct transcriptional profiles in response to multiple inducers [114] (Figure 3). NOT gates are sgRNAs which repress a gene of interest. A NOT-NOT gate can be formed by having an inducible promotor drive expression of a sgRNA, which represses another sgRNA, which in turn represses a gene of interest. NOR gates can be created by incorporating forward and reverse 5'-NGG-3' PAM sequences between the -35 and -10 regions of a promoter, allowing for targeting of either strand of DNA. The guide RNAs which target this promoter are then driven by two different inducible promoters. When either promoter is active, sgRNA will be transcribed and repress the expression of the gene of interest [114]. This NOR gate will only be active in the absence of both inducers. Other logic patterns can be formed by the combination of these techniques. These expression profiles can then be linked to the host regulatory network by designing the final sgRNA to target a native transcription factor. For example, a NOT[NOR(A,B)] gate was integrated to control the MalT expression of *E. coli* K-12 in response to arabinose and 2,4-Diacetylphloroglucinol (DAPG) inducers. MalT is a positive regulator of maltose utilization, and one consequence of repression of MalT is a decrease in LamB, the lambda phage receptor. Without the phage receptor, *E. coli* is much less susceptible to lambda phage infection. With the synthetic expression profile inserted into *E.coli*, lambda phage produced 2-3 orders of magnitude more plaques in the presence of either arabinose or DAPG than in the absence of both [114].

Intergrated circuits can be further modulated by coupling sgRNA processing to ribozymes or a type III CRISPR-Cas associated RNA endonuclease Csy4 which cleaves precursor RNAs. Csy4 recognizes a 28 nucleotide RNA sequence, cleaves the RNA, and

remains bound to the upstream cleaved fragment. These techniques allow for sgRNA expression to be coupled directly to another gene transcript such as a fluorescent marker, or for multiple sgRNA to be contained within a single transcript [120]. The control of RNA processing, for example through regulation of the expression of Csy4, can change the resulting genetic program that is produced from a single RNA transcript. For example, a single RNA transcript which contains two Csy4 recognition sequences flanking a sgRNA targeting activation of EYFP was inserted into an intron within a far-red mKate2 fluorescent marker. The resulting transcript was able to produce either red fluorescence in the absence of Csy4, or yellow fluorescence in the presence of Csy4 [120]. Transcripts can be produced with any combination of features including sgRNA, genes, RNA triplex (which stabilize and allow for translation of genes lacking poly A tails), Csy4 RNA endonuclease, introns, and ribozymes [120]. These can then be coupled to the host transcriptome, for example by expressing Csy4 on a promoter specific for a given cell-type.

**8. Synthetic Transcriptional Profiles**

Reprogramming cells into a different cell type through cellular differentiation requires precise changes to gene expression over many genes [118]. CRISPR systems afford the ability to localize these regulatory elements to specific genetic loci with pathways orthogonal to native cellular ones, and in a multiplexed way. By extending the sgRNA with an RNA domain for specific RNA binding proteins, sgRNA acts as a scaffold RNA to localize specific effector modules. In this way, a single guide RNA encodes information for both the targeting locus and which regulatory function to perform at that locus [118]. The dCas9 then acts as a master regulator for these sgRNA programs, able to perform both activation and repression from the same dCas9 protein. As already described, activation is accomplished by sgRNA recruitment of VP64 and repression is due to recruitment of KRAB. This method introduced synthetic transcriptional programs to redirect the output of the pathway of the natural violet pigment violacein between five distinct states. This highly branched metabolic pathway has several potential medical applications such as antibacterial and anticancer drugs [121]. A plasmid containing one to three sgRNA scaffolds were transformed into yeast. These sgRNA activated either VioA or VioC, or repressed VioD. Different combinations of these sgRNA redirected the flux through the violacein pathway in a controlled manner to produce related pigments such as proviolacein, violacein, prodeoxyviolacein, deoxyviolacein, or none of the aforementioned products [118]. This was detected as markedly different colored yeast cells, the presence or absence of each product being confirmed by high performance liquid chromatography.

**9. Perspectives**

CRISPR-Cas9 provides a powerful system to modify and regulate genetic sequences because it combines the efficiency and specificity of a protein based DNA-binding enzyme with the ease in target synthesis and transformation of small RNA. This lead to the explosion of CRISPR-Cas9 techniques over the past 5 years. As a result, the current limitation in perturbation experiments is generally no longer the CRISPR-Cas9. Rather, bottlenecks have become the screening and delivery methods. Even with modern ultra-high throughput screening methods such as droplet microfluidics, sequencing depth remains a bottleneck. Additionally, transformation of human cell lines with CRISPR-Cas9 is still primarily done either through transfection of the Cas9 enzyme and sgRNA directly, or with a lentiviral vector. The lentiviral vectors continue to have issues with non-specific integration and recombination between vectors which can confound results. Nevertheless, alternatives to CRISPR-Cas9 are being developed such as the Cpf1 system. It has several potential advantages over Cas9 systems. Firstly, it uses a single targeting RNA, rather

than a crRNA/tracrRNA duplex that Cas9 uses [93]. This means that Cpf1 crRNA arrays can be used *in vivo* in organisms (such as mammals) which lack the RNAse III required to process the crRNA for Cas9 as it only requires Cpf1 for crRNA processing [122]. This makes it much easier to multiplex in mammalian cells as multiple guide RNAs can be transcribed from a single promoter. Secondly, the PAM sequence for Cpf1 is T rich (5'-TTN-3') opposed to the G rich PAM in Cas9, allowing Cpf1 to directly target AT rich regulatory sequences that may not be available to Cas9. And finally Cpf1 produces staggered DNA breaks with a 4-5 nt 5' overhang, potentially allowing for ligation based repair mechanisms in addition to recombination and non-recombination end joining pathways available to Cas9 [93]. Similarly to Cas9, Cpf1 can be used for CRISPRi and CRISPRa techniques. It can be rendered catalytically inactive by the mutating the RuvC-like domain, which ends cleavage of both strands of DNA and turns it into a DNA binding protein similar to dCas9 [93]. As CRISPR methods continue to mature, it is clear that the ability to rapidly generate large libraries of genetic perturbations, or engineer a near limitless number of synthetic transcription factors, will continue to make CRISPR a staple technique in exploring and controlling gene expression in a wide range of organisms.

# 2 Interaction between Transcription Factors in Different Regulatory Clusters

The evolution of an organism is highly dependent on its adaptive landscape [123]. This landscape is the shape that occurs when mutations increase or decrease the fitness of an organism. The adaptive landscape could be smooth with a single peak representing an optimal phenotype, which the organism will tend to drift toward through the gradual accumulation of beneficial mutations. However the landscape could also be rugged, with many peaks and valleys. In these cases the directions in which an organism can move within the landscape is restricted, as drifting into a valley is less fit and therefore selected against. Organisms will tend to get stuck on a local peak and may never reach the 'true' optimal phenotype. Multiple peaks in the landscape are only possible with a specific form of epistasis referred to as reciprocal sign epistasis [124] [125]. Reciprocal sign epistasis occurs when two mutations both reduce fitness individually, but increase fitness in combination (or vice versa). However sign epistasis can vary across variable environments [26]. This provides a potential escape from these valleys.

Reciprocal sign epistasis can emerge from the structure of the regulatory system. Transcription factors in a hierarchy, or transcriptional cascade, where one transcription factor regulates the expression of another transcription factor has the potential to form reciprocal sign epistasis [25]. Because the components of the transcriptional regulatory network in *E. coli* are organized in a hierarchy, we investigated instances of reciprocal sign epistasis between different components. We perturb 4 genes located in a strongly connected component corresponding to the energy state of the cell and 4 genes in a strongly connected component corresponding to mobility. We then do the pairwise perturbations for each pair of genes across the two components. These components are considered strongly connected because there is a path for each transcription factor within them to regulate any other transcription factor in the same component, either directly or indirectly. We then record fitness measurements for all of these perturbations, using growth competition and swimming radius as two metrics for swimming to see if the natural transcriptional regulatory network showed the same reciprocal sign epistasis as a synthetic model [25]. Similar to the synthetic transcriptional cascade, we found that most sign epistasis was downstream, and we also found instances of reciprocal sign epistasis. We also confirmed that the sign epistasis was dependant on the growth media used [26]. Additionally we found that sign epistasis was not conserved across both fitness measurements. This means that variable selection pressures, like variable environments, could provide a means for the cell to escape valleys in the adaptive landscape. This also implies that as we increase the dimensionality of the system, for example by increasing the number of variable environments or selection pressures, we increase the potential evolutionary paths that the cell can take through the adaptive landscape.

## 2.1 Transcriptional regulatory network of *E. coli* consists of hierarchical strongly connected components

*With **Claire Seydoux***

We have compared the transcription factor network from RegulonDB [45] and compared it to one we have derived from bioinformatics analysis of the global feedback structures from Ecocyc database [44] and found that (1) each global regulator is strongly connected to a cluster of regulators that mutually regulate each other, (2) these clusters have hierarchical relations associated with physiological tasks (oxidation, starvation…) and (3) all regulators are downstream combinations of these few clusters (Figure 9). This is contradictory to previous depictions of the transcriptional regulatory network, which showed it as a purely hierarchical structure [39]. We also depict the sigma factors to be at the bottom output layer of this network. Here, the two most dominate sigma factors in *E. coli* rpoD and rpoS show very different positions. The first strongly connected component is the only one which regulates rpoD, while rpoS is regulated by all of the strongly connected components except the MarRA/Rob cluster. This may reflect their roles in house-keeping and stress responses respectfully. It may be appropriate to put the sigma factors at the top of the hierarchy, however this would result in the entire network becoming one strongly connected component.

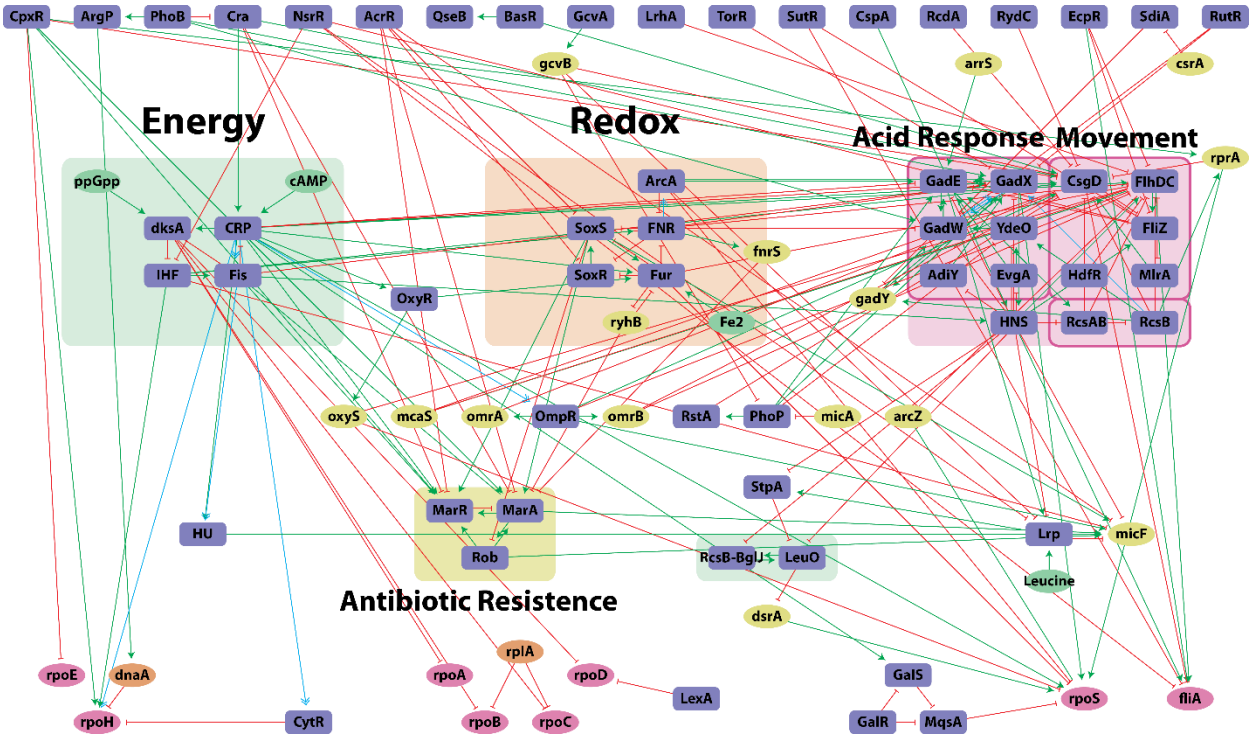



*Figure 9: Transcriptional Regulatory Network of E. coli. Transcription factors which regulate the expression level of at least one other transcription factor are shown in purple, small regulatory RNA are shown in yellow, effector molecules are shown in green, sigma factors are shown in pink and other regulatory proteins are shown in orange. Transcriptional activation is indicated by green arrows, repression by red lines and interactions which can be either as blue arrows. Strongly connected components are grouped and highlighted by a colored box. These strongly connected components correspond to physiological functions of the cell. The Energy cluster (green) contains regulators which respond directly to the energy status of the cell through ppGpp and cAMP. The Redox cluster (orange) contains transcription factors which all directly respond to the redox potential of the cell through various mechanisms. The Acid Response/Movement cluster (red) contains genes that regulate cellular mobility and respond to acid stress. The Antibiotic Resistance cluster (yellow) controls efflux pathways and provides innate defenses to antibiotics.*

## 2.2 Interactions between Strongly Connected Components

Interactions between interacting components of the regulatory network may constrain the evolutionary trajectories of cells. This can occur when the fitness effects of mutations in one component are affected by the presence of another mutation (epistasis). If the mutation is beneficial in one instance and deleterious in another, or vice versa, this is referred to as sign epistasis. When this occurs in a component involved in adaption, it constrains the possible evolutionary trajectories, avoiding the deleterious mutation [26]. It is possible for two mutations to individually be deleterious but together are beneficial, referred to as reciprocal sign epistasis. Reciprocal sign epistasis is necessary for multiple peaks within a fitness landscape, and cells can become stuck on these peaks, unable to evolve towards a higher maximum fitness because any single mutation would be selected against. However these epistatic interactions are dependent on the environment, and changes in the environmental conditions provide an escape from fitness peaks within these rugged fitness landscapes [26]. If changes to environment provide additional evolutionary trajectories, it is possible that changes to selection pressures might as well. It is not clear if epistatic traits between genes would hold across multiple fitness measurements, such as growth or mobility, and if not could competing selection pressures overcome restrictions from reciprocal sign epistasis?

Within the transcriptional regulatory network of *E. coli*, there are multiple strongly connected components. These are groups of transcription factors which are able to directly or indirectly regulate all other transcription factors within that group. The members of a strongly connected component appear to be functionally related. For example FNR, ArcA, Fur, SoxR, and SoxS are all in a single strongly connected component and all 5 transcription factors directly sense the oxidative state of the cell, which modulates their functionality. Sign epistasis has recently been shown within transcriptional cascades [25]. This motivated us to explore how perturbations in one strongly connected component would influence phenotypes that structurally we associate with another strongly connected component, and how the hierarchy between strongly connected components influences the epistasis of between genes in different strongly connected components. To this end, we chose to study the influence between the strongly connected component associated with the energy state of the cell and the strongly connected component associated with cellular mobility. In this context, we can quantify the proportion of upstream sign epistasis (that is within the energy cluster), downstream sign epistasis (within the mobility cluster), and reciprocal sign epistasis. We can also see if there is a tendency for sign epistasis to be coupled between fitness measurements, and if so, correspond to the regulatory network structure?

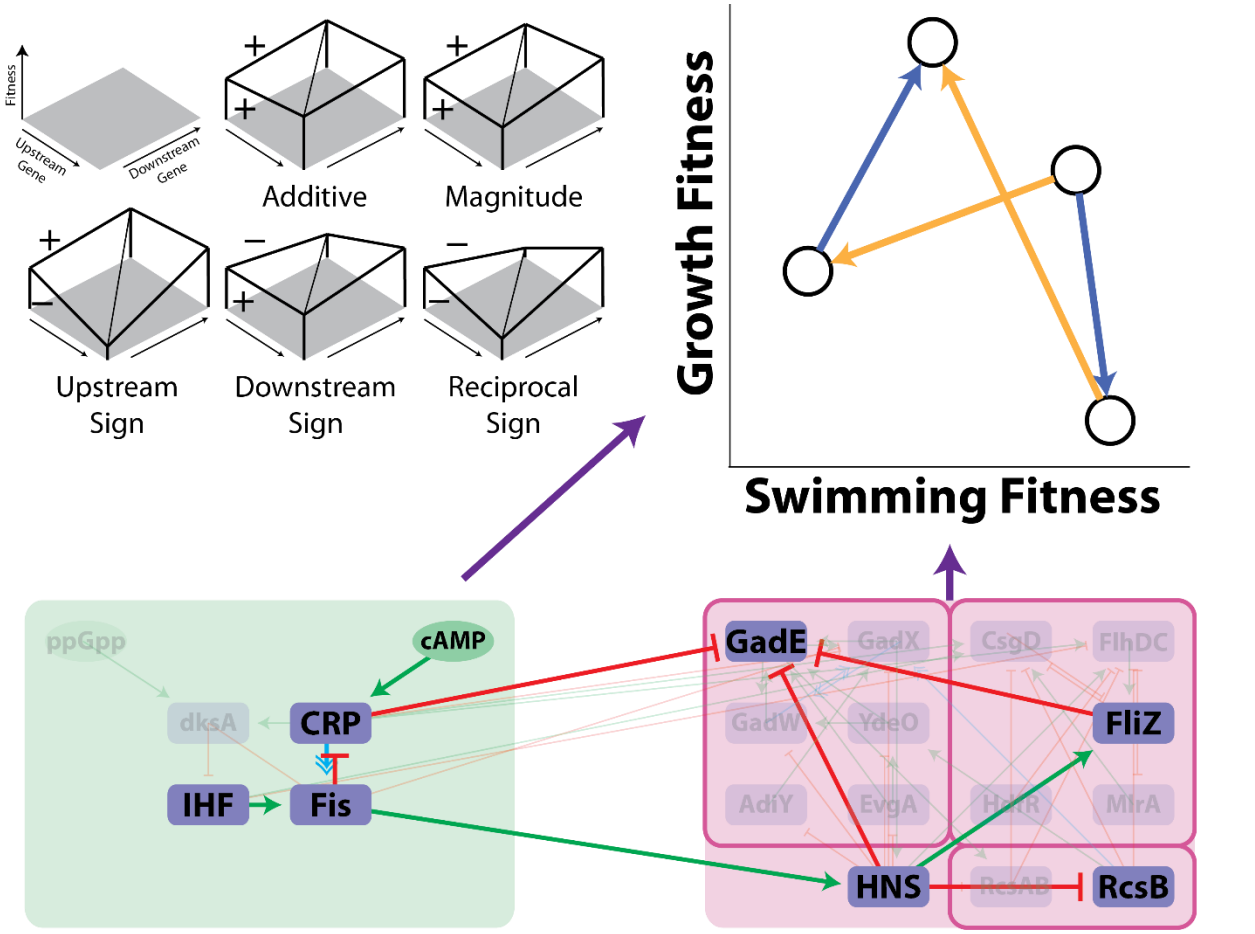



*Figure 10: Fitness landscapes between two genes can be flat (additive, no epistasis) or bent (epistasis). If the bend simply increases or decreases the slope of the landscape, but does not change its direction it is referred to as magnitude epistasis. However in some cases the effect of a gene will decrease fitness in one instance and increase fitness in another (or vice versa), referred to as sign epistasis. This can occur for either or both genes. If the direction of both genes change, it is specifically referred to as reciprocal sign epistasis and causes a trench to form in the fitness landscape. The trench is difficult to evolve past as each initial mutation will cause a fitness decrease and be selected for. Hierarchy in transcriptional cascades has previously been shown to cause sign epistasis, therefore we look for sign epistasis between two hierarchical clusters of transcriptional regulators in E. coli. We measure fitness by two different fitness metrics, one associated primarily (though not exclusively) with each cluster. We then look for fitness trajectories in both fitness metrics for single and double perturbations of genes within these clusters. The pair of genes may show reciprocal sign epistasis in one fitness but not the other (above: growth reciprocal sign epistasis, swimming magnitude epistasis).*

The energy state component contains the transcription factors IHF, CRP, and Fis, and directly responds to the cofactors cAMP which is a signal for the quality of carbon source in the growth media, and ppGpp which is an alarmone signaling stalls in protein synthesis, due to starvation or stress conditions triggering the stringent response. This is the first strongly connected component in the transcriptional regulatory network and is upstream of most other transcription factors as well as all the additional strongly connected components. The component associated with cellular mobility contains the master regulator for the flagella pathway FliZ, as well as CsgD and FlhDC, which act as a toggle switch between twitching and swimming mobility. In addition to mobility, this component also has transcription factors responsible for the acid response pathway such as GadE, GadX and GadW, as well as Response regulators RcsB and the regulator for horizontally acquired genes HNS.

We use a growth fitness and a swimming fitness as two easily quantifiable measurements related to the IHF, CRP, Fis and the FliZ, GadE, HNS, RcsB strongly connected components respectively. The growth fitness is quantified by performing a competition assay between a non-perturbed reference strain and a genetically perturbed strain, in which one or two genes are repressed using the CRIPSR-dCas9 system. The relative proportion of each strain is measured using a fluorescent marker (either EGFP, mCherry, or mCerulean) and the Optical Denstity (OD) of the entire culture is recorded. We are interested in the change in the relative fluorescence between the two strains as a function of the change in OD. This is normalized against a competition assay containing two reference strains to account for differences in maturation or fitness costs of the different fluorescent markers, and forces the slope of the reference strain to zero. Perturbed strains which are out competing the reference strain will have a positive slope and strains which are being out competed will have a negative slope. This slope can be used as an exponent to determine the fitness, where $f_{growth}$ = 2^Slope. This sets the reference strain fitness to 1. Swimming fitness is quantified by the distance that strains can swim in a soft agar plate, where the percent agar is 0.3% or lower. A cell culture is spotted onto the center of a soft agar plate and incubated for 16-24 hours. This can also be done in competition with a reference strain using two different fluorescent markers. With a perturbed strain tagged with EGFP and a reference strain tagged with mCherry, a competition swimming assay should result in two observable circles. Firstly, an interior yellow circle will be the distance covered by both strains. Second, either a green or red outer ring will indicate which strain out competes the other. If the ring is green, the perturbed strain was able to swim farther than the reference strain and if the ring is red, the opposite is true. The radius of these rings can be determined by image analysis and again normalized by the swimming of a reference strain against another reference strain, to account for differences caused by the fluorescent markers.

To perform these experiments we modified pCRRNA vector supplied by Lun Cui and David Bikard from Institut Pasteur. We inserted either EGFP, mCherry, or mCerulean between the crRNA array and the origin of replication. The fluorescent marker was under the control of a strong constitutive promoter (J23119) and terminated with a bidirectional rho independent terminator (B0014). The fluorescent marker was the opposite direction to the crRNA promoter, such that there should be no transcriptional read-through, nor should transcription of the marker reduce crRNA expression due to supercoiling effects. Single and double spacers were then inserted into the crRNA array using golden gate assembly to generate vectors. These knocked down transcription of either crp, fis, ihfA, ihfB, fliZ, gadE, hns, or rcsB individually or in combinations of one of the first four with one of the last four aforementioned genes. This resulted in 24 pCRRNA vectors. We also inserted non-targeting spacers into fluorescent pCRRNA vectors to act as reference strains. The fluorescent pCRRNA vectors were transformed into LC-E24 :: dcas9 2tetO HK022 attB and MG1655 pdCas9 host strains which have dCas9 either integrated into the chromosome or on a separate plasmid respectively. All vectors were sequenced to ensure the corrected targeting spacer was inserted, and all strains were tested with qRT-PCR to ensure that the CRISPR-dCas9 was repressing the gene of interest.

## 2.3 Growth Competition and Swimming Competition fitness

*With **Marine Lombard***

***Genome integrated dCas9 demonstrated inhibited swimming from a cloning artifact.***

Initially we performed growth and swimming assays by inserting our fluorescent pCRRNA vectors into LC-E24 :: dcas9 2tetO HK022 attB. We performed Growth assays on LB media, and M63 media containing either Glucose or Lactose as a Carbon Source. In LB media, all the perturbations reduced the fitness of strain in competition assays. In our earlier tests with pCKDL vectors, the single crp, fis, and hns perturbations did not show significant differences in maximum growth rate to the reference strain. We saw much stronger negative fitness measurements when crp, fis, and hns were perturbed in competition assays. This is expected due to the much higher sensitivity of competition assays compared to growth curves. The biological replicates are separated on the Y axis, which indicates variations in the starting ratios of the perturbation strains and the reference strains on separate days. However this does not change the direction of the slope of the curves. We found that the competition fitness was media dependent. When we changed to a defined media M63, certain perturbations were advantageous. This further depended on the carbon source available. In M63 with Glucose as a carbon source, gadE perturbation became advantageous. We also observed sign epistasis between some pairs of genes, specifically crp-fliZ, ihfB-hns, and ihfB-rcsB, because these pairs of perturbations had a fitness advantage although all of the individual perturbations caused a fitness disadvantage. We also saw higher than expected fitness advantage for fis-gadE, and ihfB-gadE. In M63 with Glucose we do have cases when the slope was not reproduced in separated experiments, namely crp-gadE and fis-rcsB showed different fitness in different replicates. This could be due to individual mutants escaping the CRISPR-dCas9 system. When the carbon source was changed to lactose, we found even more perturbations which gave a fitness advantage. Here, 12 strains had a fitness advantage over the reference strain. Additionally, reproducibility was decreased, with 5 strains showing inconsistent phenotypes. Overall, the diversity of the growth competition phenotypes increases as we increased competitive stress by decreasing the richness of the growth media.

When we examined swimming phenotypes with these strains, we found that the reference strain with mCherry was not reproducibly swimming. Additionally, strains seemed to swim from one or more blooms from the centre culture where the plate was inoculated with the cell mixture. We tested MG1655 and LC-E24 :: dcas9 2tetO HK022 attB and found that even without any pCRRNA vector, and LC-E24 :: dcas9 2tetO HK022 attB showed this same inconsistent and asymmetrical swimming pattern. We transformed pdCas9 into MG1655 and found that it showed a symmetrical and reproducible swimming pattern similar to that of MG1655, although with a decreased radius. We also tested a cloning strain (Top10) and found that like LC-E24 it had difficulty swimming. We concluded that the observed swimming phenotype in LC-E24 was likely a cloning artifact.

***Growth Fitness by competition assays using pdCas9 reference strain shows sign epistasis.***

We therefore repeated our experiments using MG1655 pdCas9 as a host strain (Figure 11). We found similar growth results to our LC-E24 strain experiments however they did have some differences. Notably, the results were less reproducible in MG1655 pdCas9. This could be due to the increased burden of an additional plasmid pdCas9 [126]. Not only does the additional plasmid and antibiotic increase the metabolic load on the cells, but the additional dCas9 produced increases the likelihood of off-target

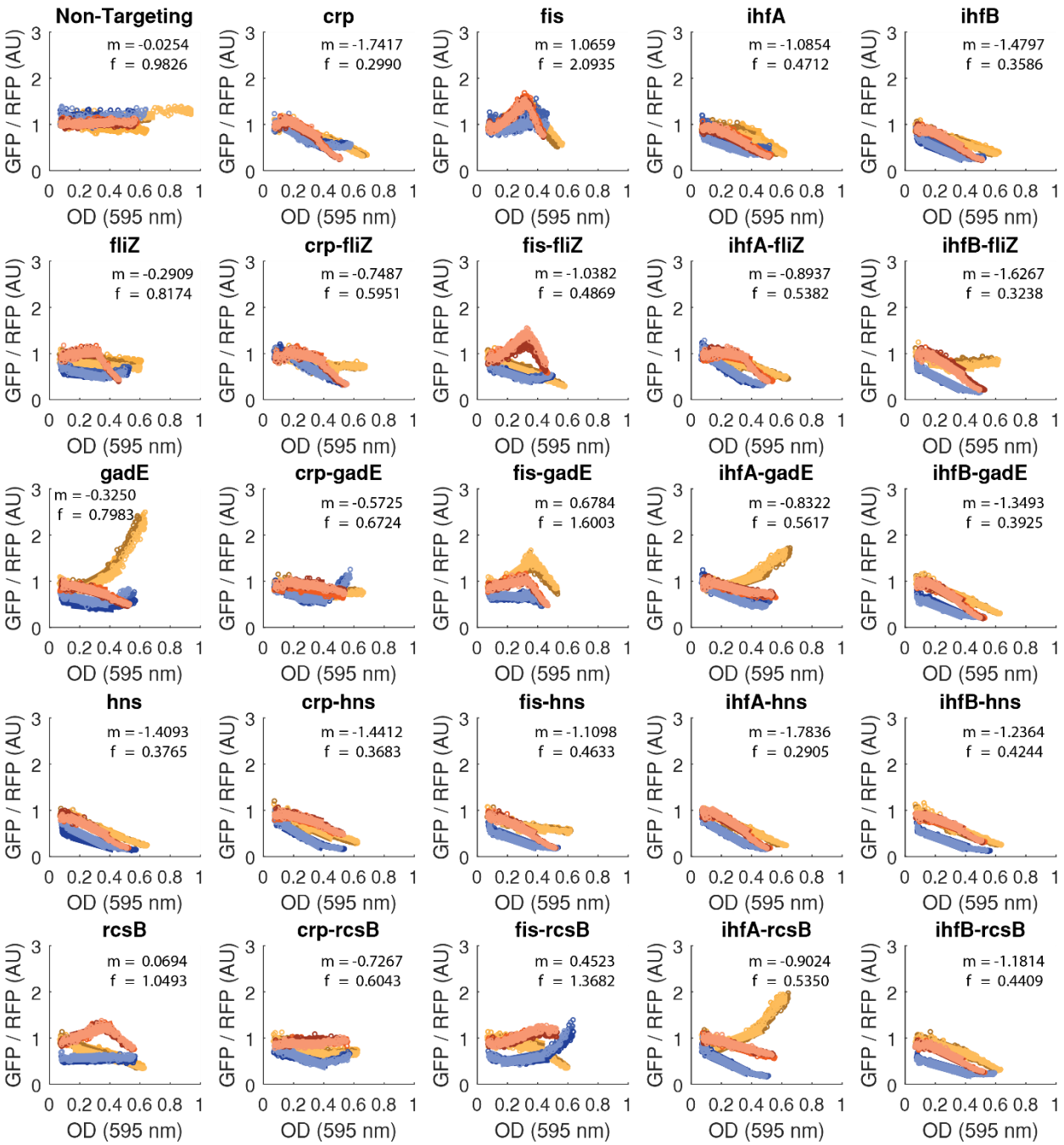



*Figure 11: Competition assays between MG1655 pdCas9 pCRRNA :: EGFP and MG1655 pdCas9 pCRRNA :: mCherry in LB media. Strains harboring pCRRNA :: EGFP carried on of 24 targeting spacers or a non-targeting control (indicated at the top of the plots). The strain harboring pCRRNA :: mCherry contained a non-targeting control spacer and used as a reference strain. All EGFP strains were grown in competition with the reference strain, at a starting ratio of 1:1 made from dilution of 1/100 from overnight cultures. The change in the ratio of GFP to RFP signal with the change in optical density (OD) indicates the relative change in proportion of each strain as cultures grew. A slope was fitted to the data between the 25-75% of OD to avoid changes in the lag phase (which may be instrument noise) and stationary phase. A slope (m) of 0 indicates no change in relative proportions of each strain. A fitness score (f) was determined for each strain by computing 2^m. Data represents 3 biological replicates with 3 technical replicates each.*

effects. Additionally, the uniform over performance of perturbations in M63 + Lactose, along with a few extremely high fitness scores ($4.94 * 10^4$ in the most extreme case) indicates a systematic artifact in those samples. We found 4 incidents which indicated sign epistasis in LB, three downstream and one upstream, and 3 of the incidents of sign epistasis involved pairs with FliZ perturbations. We had expected FliZ to have a large impact on swimming mobility, but did not expect much effect in competition assays as it is not highly expressed in our qRT-PCRs and is abundant after exponential growth [44]. However FliZ is known to act as an antagonist to sigma-S and therefore its disruption maybe effecting the transition from exponential to stationary phase. In M63 media with 0.4% Glucose, we found 6 incidents of sign epistasis (3 reciprocal, 2 downstream, 1 upstream). Here again, two genes showed sign epistasis in 3 of their 4 pairwise perturbations. These were Fis and GadE. In cases were sign epistasis was not reciprocal, both genes were the gene in the pairwise strain to show sign epistasis (upstream for Fis and downstream for GadE). Changing Glucose for Lactose resulted in 7 cases of sign epistasis (3 reciprocal, 3 downstream, 1 upstream). Again there was one gene (in this case RcsB) which showed epistasis in 3 out of 4 interactions, and was consistently reciprocal or downstream. While these results shouldn't be over interpreted, they are consistent with the previous finding that downstream sign epistasis is more common in transcriptional cascades than upstream sign epistasis [25], and specific regulators tend to consistently demonstrate sign epistasis more than other regulators in terms of growth fitness.

***Sign epistasis in swimming fitness with pdCas9 reference strain***

To examine if these trends held with another fitness measurement, we performed swimming experiments with MG1655 pdCas9 pCRRNA :: EGFP. We imaged our plates with a fixed camera and used edge detection in matlab to locate the boundaries of swimming (Figure 12). Fortunately, the reference strains generally had an expected swimming phenotype. However we found that the swimming phenotype was dependent on both the aTc concentration and on the % agar in the soft agar plates. We also found that the blooming phenotype still occurred in many cases. In these cases we took the radius of the cells not including any blooms. Again we found incidents of sign epistasis (Figure 13); in LB we found 5 cases (3 downstream, 1 upstream, 1 reciprocal), in M63 with 0.4% Glucose we found only 1 case of reciprocal sign epistasis, and in M63 with 0.4% Lactose we found 6 cases (3 downstream, 3 upstream). While overall we found more downstream sign epistasis than upstream, it was not as strong as in growth cultures. We only found 2 strains that showed sign epistasis in 3 out of 4 of their gene pairs, fis and rcsB grown in M63 with Lactose. Unfortunately, in this case these perturbations were not consistent, with 2 downstream and 1 upstream sign epistasis for each. We found that outside of LB media, perturbation of FliZ generally suppressed any swimming phenotype. This is consistent with FliZ regulating genes within the mobility pathway [44].

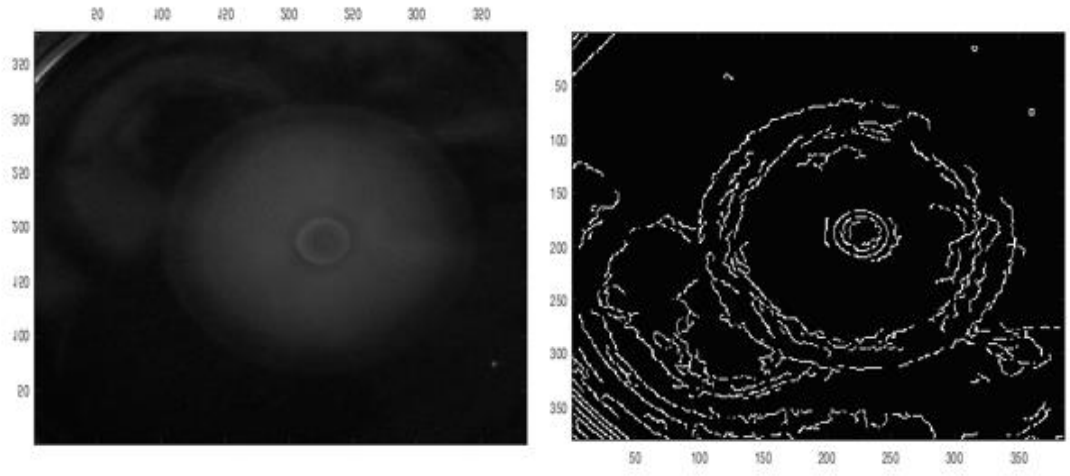


*Figure 12: Detection of swimming radius using matlab edge detection. Soft agar plates were imaged using a fixed camera (left). Images were then analysed in matlab using edge detection to determine the radius of the swimming bacteria (right).*

It is possible that the difficulty in producing consistent results in both the swimming and growth assays is due to the high cellular burden being imposed on the cell. We see a consistent reduction in both growth rate and swimming mobility with the induction of pdCas9 alone, which gets progressively worse with the addition of pCRRNA. Excessive expression of additional proteins has long been known to adversely affect the growth rate and health of the cell, and can lead to break down and loss of rRNAs, ribosomes, and protein synthesis capacity [127]. While we can adjust the expression of dCas9 by titrating anhydrotetracycline (aTc), our fluorescent markers are on strong constitutive promoters for visualization during swimming assays. Additionally, none of our proteins have been codon optimized for *E. coli*, which when paired with high expression levels could lead to depletion of rare tRNAs or reduced mRNA stability [128]. These costs may have a strong impact on the cellular decision to swim, as the flagella is a very costly structure to create and maintain for the cell [129] [130]. We may have more reproducible and reliable results by running dCas9, crRNA, and tracrRNA from a single plasmid, without fluorescence. This will prevent competition assays from being performed, but they could be replaced with growth curves.

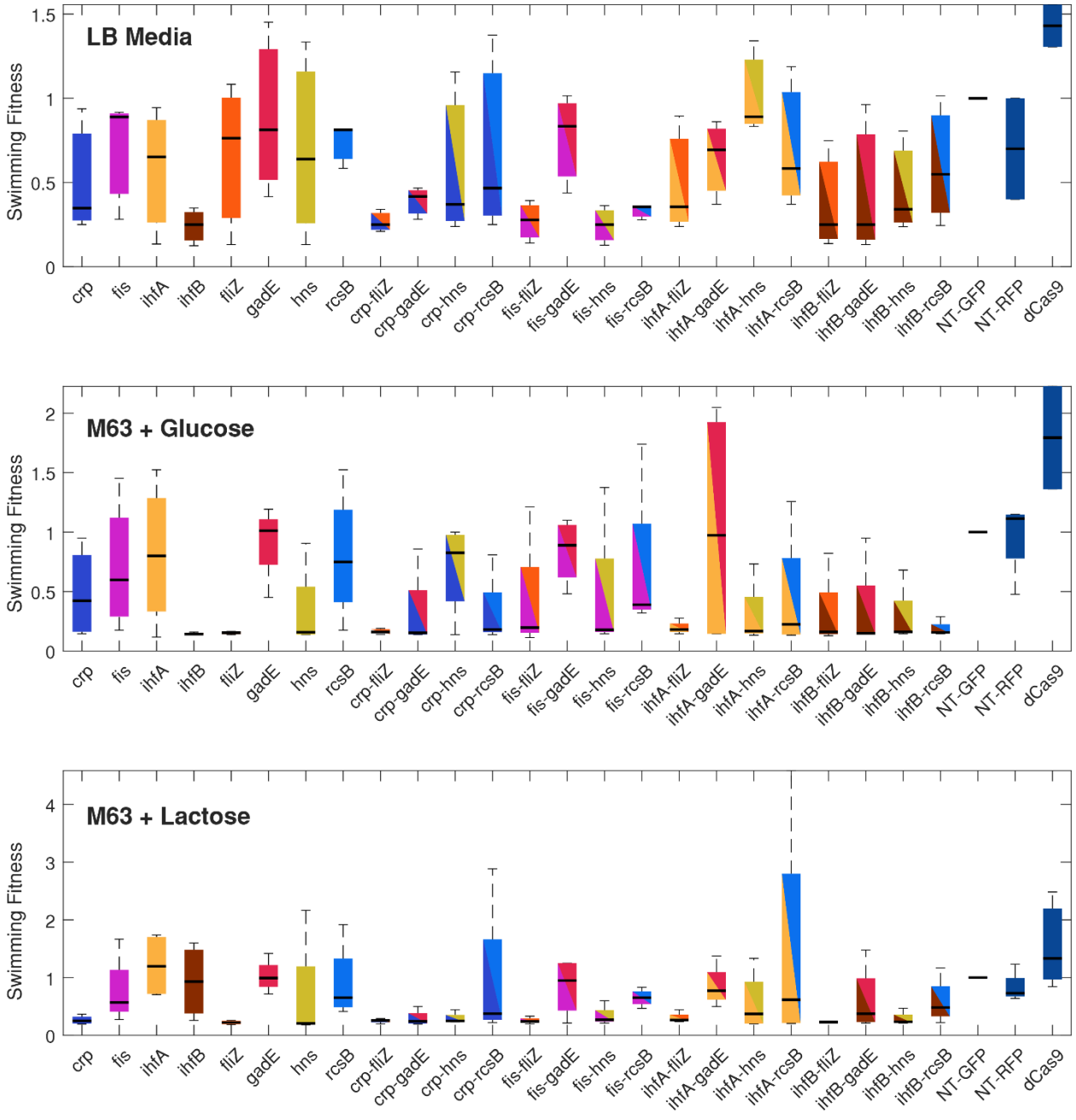



*Figure 13: Swimming mobility of single and double knockdown MG1655 pdCas9 pCRRNA :: EGFP strains. Strains were spotted onto growth media containing 0.25% agar. The radius of swimming bacteria was recorded after 16 or 20 hours for LB and M63 respectively. The radius of each strain was normalized to the radius of the reference strain with a non-targeting (NT) spacer. Data represents 3-4 independent replicates for each strain.*

## 2.4 Fitness Landscape of interactions between 'Energy' and 'Mobility' Regulatory Clusters

***Shapes on the fitness trajectories represent different types of epistasis***

For each pair of perturbations we plotted the growth and swimming fitness measurements for the reference strain, the individual perturbations and the combination of perturbations for LB media, M63 with Glucose and M6 with Lactose. Since we have normalized the fitness of the reference strain to 1, the predicted fitness of the double perturbation strain (AB) with no epistasis is fAB = fref * fA * fB. This prevents negative fitness from being possible, as it would be impossible to have a negative swimming diameter or a negative growth rate. With this we calculated all interactions that would have an expected fitness with no epistasis. We plotted the fitness measurements for each pair of genes against each other in each media (Figure 14). The direction and the shape of the trajectories indicates the type of epistasis between the two genes. If the trajectories for a given gene are in the opposite direction in relation to one or two axis, they indicate sign epistasis for that gene in the corresponding fitness. If the arrows have a different length but the same direction, they indicate magnitude epistasis. We identified 6 basic shapes within our data. Parallelogram, Obtuse, Acute, Triangles, Concave, and Hourglasses. Parallelograms always indicate no epistasis. Obtuse quadrilaterals indicate magnitude epistasis. Acute quadrilaterals will indicate sign epistasis on the two side edges with acute angles with the long edge as long as neither edge is parallel with one of the axis, these sometimes occur as Acute Trapezoids (for example CRP + RcsB in LB). Triangles have two strains with the same fitness measurements resulting in one edge length near zero, when it is a single and double mutant it indicates that one perturbation is completely masking the influence of the other perturbation. When it is the reference strain and a single mutant, it indicates that one perturbation has no influence without the second perturbation. Concave shapes indicate reciprocal sign epistasis for at least one fitness measurement, and depending on the orientation, possibly both. Finally, Hourglass shapes have fitness that cross over. The base of the hourglass will have sign epistasis in both fitness metrics unless the parallel with one axis in which case it will only be sign epistasis in one fitness. Depending on the rotation of the hourglass, it may also have sign epistasis for the other gene, resulting in reciprocal sign epistasis, however it can only have reciprocal sign epistasis for one fitness. It is important to note that the shape alone does not determine the epistasis, but the rotation of the shape must also be taken into account.

| | | **CRP** | | | **Fis** | | | **IhfA** | | | **IhfB** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *LB* | *Glucose* | *Lactose* | *LB* | *Glucose* | *Lactose* | *LB* | *Glucose* | *Lactose* | *LB* | *Glucose* | *Lactose* |
| **Fliz** | *Growth* | **S (Dn)** | **M** | **M** | **S (Up)** | **R** | **S (Up)** | **S (Dn)** | **M** | **M** | **A** | **A** | **M** |
| | *Swim* | **A** | **A** | **M** | **M** | **M** | **M** | **M** | **A** | **A** | **A** | **M** | **A** |
| **GadE** | *Growth* | **S (Dn)** | **S (Dn)** | **M** | **M** | **R** | **S (Dn)** | **M** | **A** | **M** | **A** | **S (Dn)** | **R** |
| | *Swim* | **S (Dn)** | **M** | **A** | **S (Up)** | **M** | **S (Dn)** | **M** | **M** | **S (Up)** | **A** | **A** | **M** |
| **HNS** | *Growth* | **M** | **M** | **R** | **M** | **M** | **M** | **M** | **A** | **M** | **M** | **R** | **M** |
| | *Swim* | **M** | **R** | **M** | **M** | **A** | **S (Up)** | **R** | **A** | **M** | **M** | **M** | **A** |
| **RcsB** | *Growth* | **M** | **A** | **S (Dn)** | **M** | **S (Up)** | **M** | **A** | **M** | **S (Dn)** | **A** | **A** | **R** |
| | *Swim* | **S (Dn)** | **M** | **S (Dn)** | **M** | **A** | **S (Dn)** | **M** | **M** | **S (Up)** | **S (Dn)** | **M** | **M** |

*Table 1: Summary of epistatic interaction in swimming and growth fitness metrics for LB and M63 media. Columns represent upstream knock-downs while rows represent downstream knock-downs using CRISPR-dCas9 system. Interactions include: Additive – No epistasis (**A**), Magnitude (**M**), Downstream Sign (**S (Dn)**), Upstream Sign (**S (Up)**), and Reciprocal Sign (**R**). There are 6 measurements for each pair of knock-downs, three media (LB, M63 + Glucose, M63 + Lactose) and 2 Fitness metrics (Growth and Swimming).*

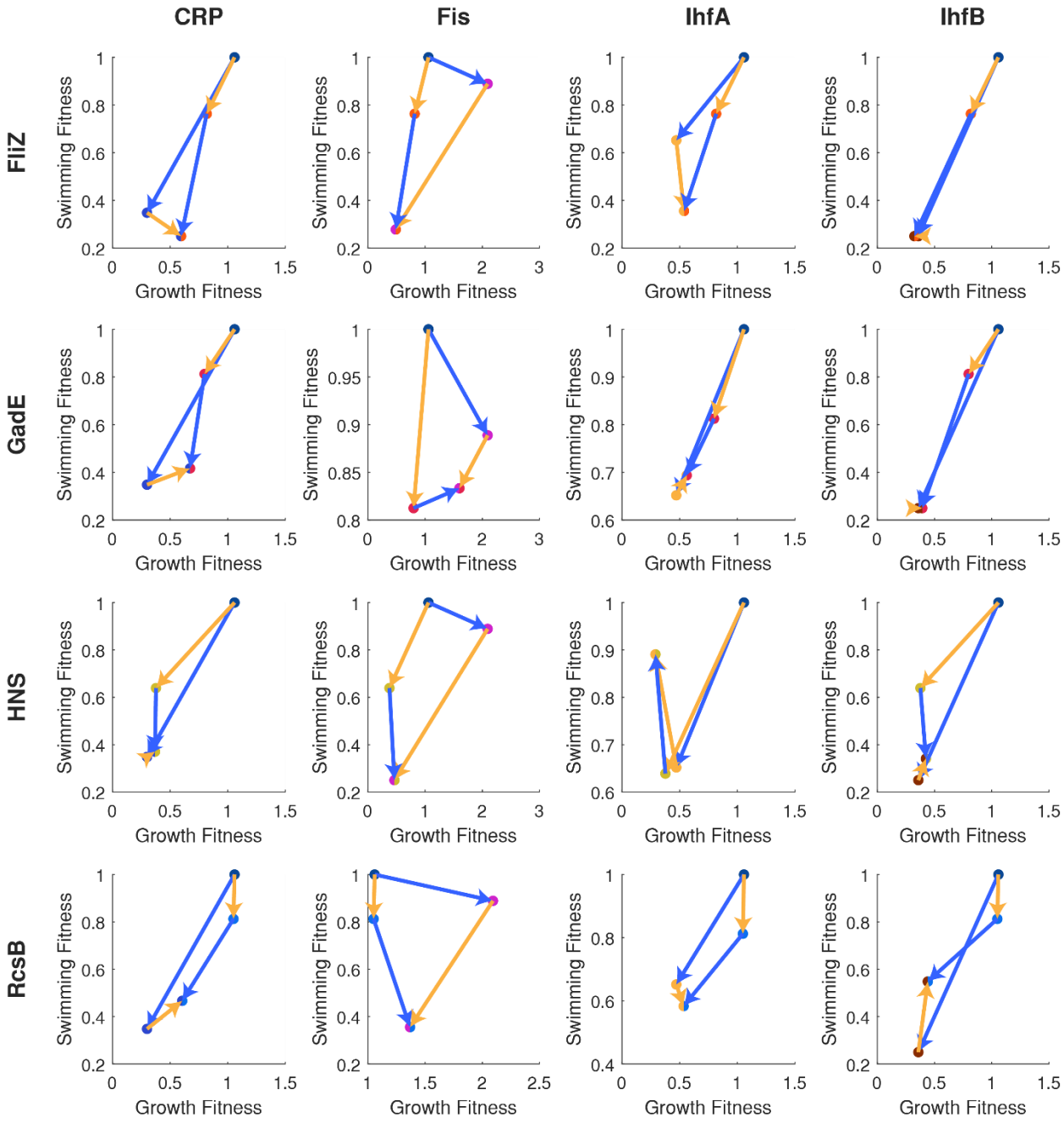



*Figure 14: Fitness trajectories for MG1655 pdCas9 pCRRNA :: EGFP strains in LB media. Fitness measurements for each crRNA spacer were normalized to the reference strain (non-targeting spacer) such that the reference strain had a fitness of 1. Single and double knock-down mutants were then plotted for each combination of strains. Blue arrows indicate the fitness trajectory with the addition of the upstream (columns) gene while yellow arrows indicate the trajectory with the downstream (rows) genes. Trajectories with opposite directions in one or two axis represent sign epistasis for the respective fitness.*

***Sign epistasis is dependent on both environment and selection pressure***

We found that, similar to previously reported studies [26], the epistasis depended on the environment of the cells, and changing the growth media changed the observed epistasis (Table 1). We also had many instances where fitness trajectories moved along the diagonal. This implies that the fitness metrics may be coupled, although it warrants further investigation. Additionally, while we did find reciprocal epistasis, we didn't find any instances where perturbations had reciprocal epistasis in both fitness metrics. We did find some examples where sign epistasis was found in both fitness metrics but in all four cases it was downstream sign epistasis. This implies that multiple selective pressures, or alternating selective pressures, could lead to additional evolutionary paths with which to avoid reciprocal epistasis. For example, a colony of bacteria face fierce competitive fitness during growth, with cells that are able to divide faster outcompeting cells which divide slower. If these slower dividing cells have an advantage in other areas, such as mobility, they may reach nutrient sources that faster dividing cells cannot. Further mutations may occur at these new nutrient sources which then decrease mobility but increase cell division, again providing a competitive advantage in a new environment. In the cases of Concave and Hourglass shapes in the fitness landscape, this could allow cells to reach new spaces on the Pareto front which would otherwise not be accessible through only one fitness selection due to reciprocal sign epistasis. The Pareto front is a frontier in the fitness landscape in which no other phenotypes exist which are better at all fitnesses [131]. Natural selection is thought to push phenotypes towards this frontier. Gene expression in *E. coli* has previously been demonstrated to fall along a line with one end repressing growth and the other representing stress, using this Pareto technique [53].

## 2.5 Methodology

*Creation of vectors*

pCRRNA mCherry and pCRRNA EGFP were made by Gibson Assembly. PCR was performed on pMD019 (pGFP) and pMD024 (pmCherry) with primers oMD546 and oMD547, and on pCRRNA with primers oMD545 and oMD548. PCR products were purified with PCR clean up kit from Macherey-Nagel. Equal molar concentrations of PCR product were mixed (one from either pMD019 or pMD024 and one from pCRRNA) and 5 µL of mix was added to 15 µL of Gibson Master mix to create pCRRNA Green and pCRRNA Red respectively. Golden Gate Assembly was used to replace the TrrnB terminator from pMD019 and pMD024 with B0014 as the TrrnB terminator contained 2 BsaI sites. PCR was done on pCRRNA Green and pCRRNA Red with primers oMD609 and oMD610, and on pCKDL with primers oMD607 and oMD608. PCR products where joined with Golden Gate Assembly using BsmBI enzyme to create pCRRNA EGFP and pCRRNA mCherry. pCRRNA mCerulean was created by Gibson Assembly. PCR was performed on pMD027 (pCerulean) with primers oMD704 and oMD705, and on pCRRNA mCherry with primers oMD706 and oMD707). PCR products were purified with PCR clean up kit from Macherey-Nagel. Equal molar concentrations of PCR product were mixed and 5 µL of mix was added to 15 µL of Gibson Master mix. All vectors were sequenced by GATC Biotech prior to use.

*Growth Conditions of Cultures*

The host strain for all pCRRNA vectors is MG1655 with the pdCas9 vector from Stanley Qi (provided by Lun Cui and David Bikard). Glyercol stocks of each culture were streaked onto individual Lysogeny Broth (LB) agar plates containing 34 µg/mL of Chloramphenicol and 50 µg/mL of Kanamycin. Single colonies were inoculated into 2 mL of selected media (either LB or M63 supplemented with 0.4% Glucose or Lactose) containing 34 µg/mL Chloramphenicol and 50 µg/mL Kanamycin. Cultures were placed in a 37$^{o}$C incubator for either overnight for 16 hours for LB cultures or for 24 hours for M63 cultures.

*Growth Competition Assays*

Assays were performed by diluting 220 µL pCRRNA mCherry NT pre-culture into 22 mL of selected media containing 34 µg/mL Chloramphenicol, 50 µg/mL Kanamycin, and 250 ng/mL anhydrotetracycline. A Greiner, 96 Well, PS, F-Bottom, µCLEAR, Black microplate was filled with 198 µL of diluted culture per well. For each pCRRNA EGFP knockdown vector, 2 µL of pre-culture was inoculated into 3 individual wells. For the Non-targeting Control strain, 2 µL of pre-culture was inoculated into 15 individual wells. A volume of 60 µL of mineral oil was added to each well of the microplate. The Absorbance at 595 nm, as well as fluorescence at 480/510 nm and 580/610 nm (excitation/emission) were recorded every 10 minutes for 20 hours with a SpectraMax i3x. Microplates were incubated at 37$^{o}$C and shook for 90 seconds before and after each measurement. To determine the fitness measurement of each knockdown, the ratio of the green and red fluorescence of the strain was divided by the median ratio of the green and red fluorescence of the 12 additional wells for the non-targeting control. Fitness measurements for each knockdown were made for LB media, and M63 media containing 0.4% of either glucose or lactose.

*Swimming Fitness Assays*

For each knockdown, 10 µL of pre-culture was spotted into the middle of 15 mL of selected media soft agar (0.3%) plates. Plates where then incubated at 37$^{o}$C for either 16 hours for LB plates or 24 hours

for M63 plates. Plates where then imaged using a USB Camera and python viewer. An image of a non-inoculated plate was subtracted from each image. The fitness was determined by the ratio of the swimming area of each knockdown strain to that of the non-targeting control strain.

## 2.6 Supplementary Figures

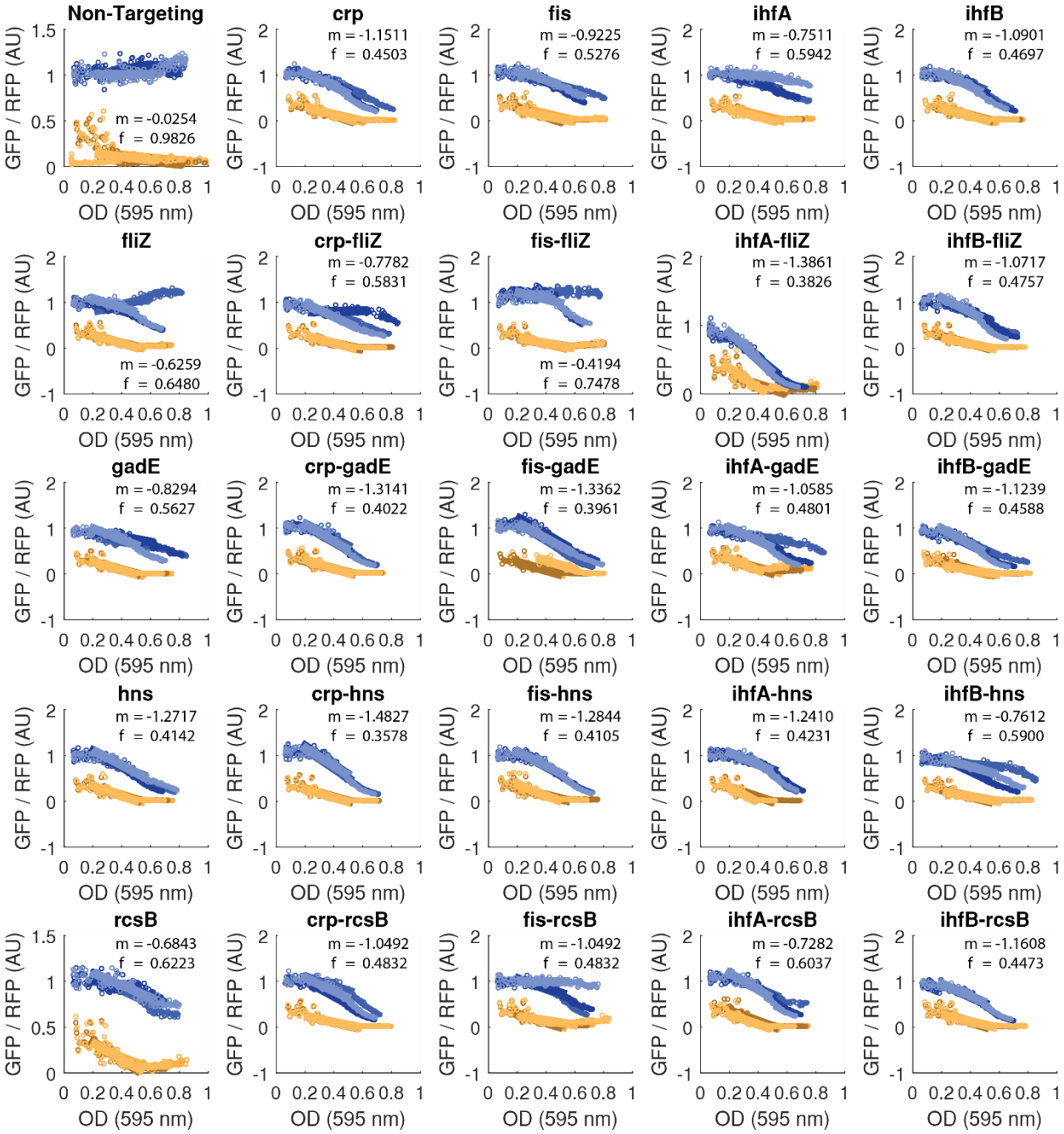



*Figure 15: Competition assays between LC-E24 :: dcas9 2tetO HK022 attB pCRRNA :: EGFP and LC-E24 :: dcas9 2tetO HK022 attB pCRRNA :: mCherry in LB media. Strains harboring pCRRNA :: EGFP carried on of 24 targeting spacers or a non-targeting control (indicated at the top of the plots). The strain harboring pCRRNA :: mCherry contained a non-targeting control spacer and used as a reference strain. All EGFP strains were grown in competition with the reference strain, at a starting ratio of 1:1 made from dilution of 1/100 from overnight cultures. The change in the ratio of GFP to RFP signal with the change in optical density (OD) indicates the relative change in proportion of each strain as cultures grew. A slope was fitted to the data between the 25-75% of OD to avoid changes in the lag phase (which may be instrument noise) and stationary phase. A slope (m) of 0 indicates no change in relative proportions of each strain. A fitness score (f) was determined for each strain by computing 2^m. Data represents 2 biological replicates with 3 technical replicates each.*

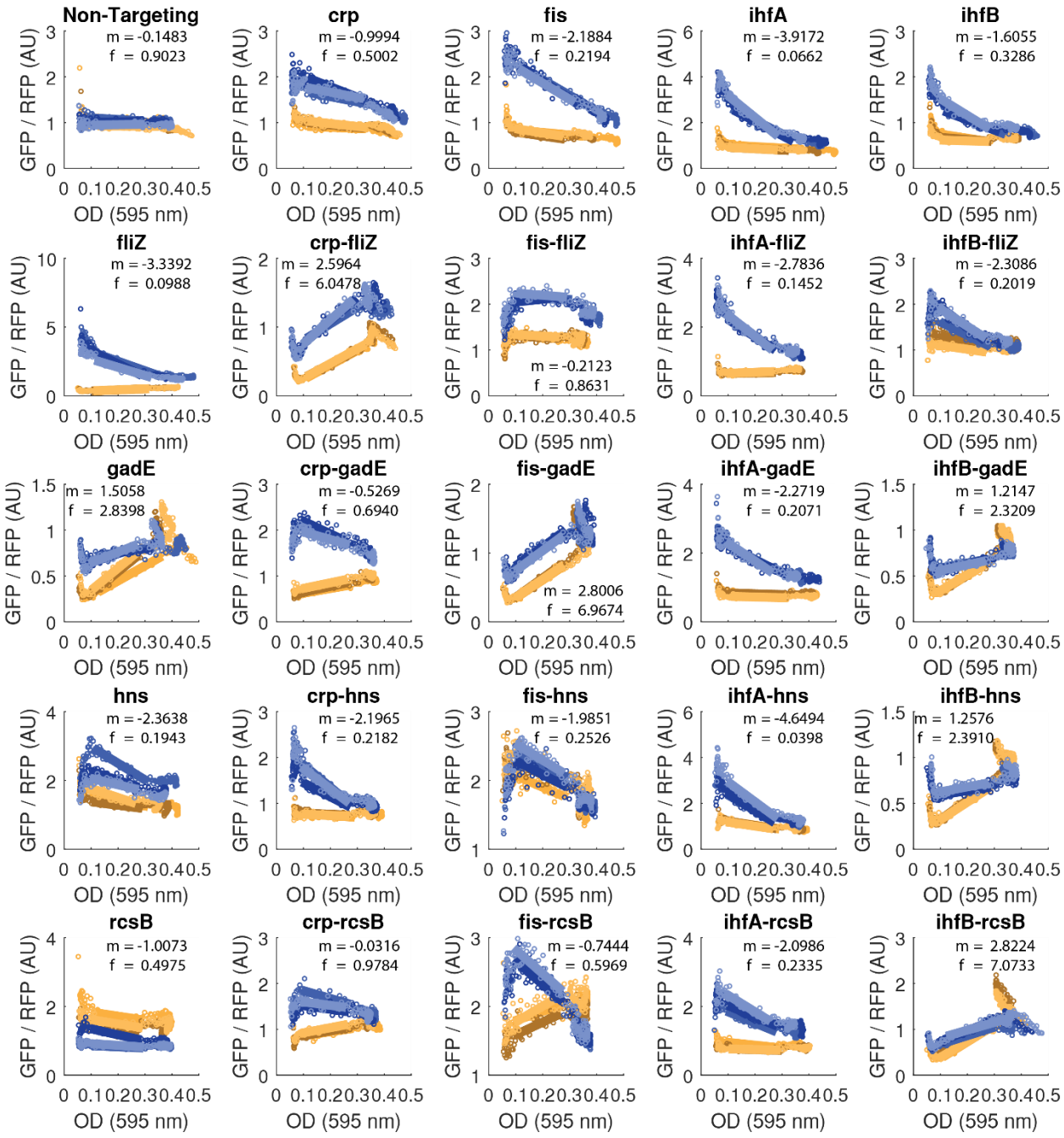



*Figure 16: Competition assays between LC-E24 :: dcas9 2tetO HK022 attB pCRRNA :: EGFP and LC-E24 :: dcas9 2tetO HK022 attB pCRRNA :: mCherry in M9 media with 0.4% Glucose. Strains harboring pCRRNA :: EGFP carried on of 24 targeting spacers or a non-targeting control (indicated at the top of the plots). The strain harboring pCRRNA :: mCherry contained a non-targeting control spacer and used as a reference strain. All EGFP strains were grown in competition with the reference strain, at a starting ratio of 1:1 made from dilution of 1/100 from overnight cultures. The change in the ratio of GFP to RFP signal with the change in optical density (OD) indicates the relative change in proportion of each strain as cultures grew. A slope was fitted to the data between the 25-75% of OD to avoid changes in the lag phase (which may be instrument noise) and stationary phase. A slope (m) of 0 indicates no change in relative proportions of each strain. A fitness score (f) was determined for each strain by computing 2^m. Data represents 2 biological replicates with 3 technical replicates each.*

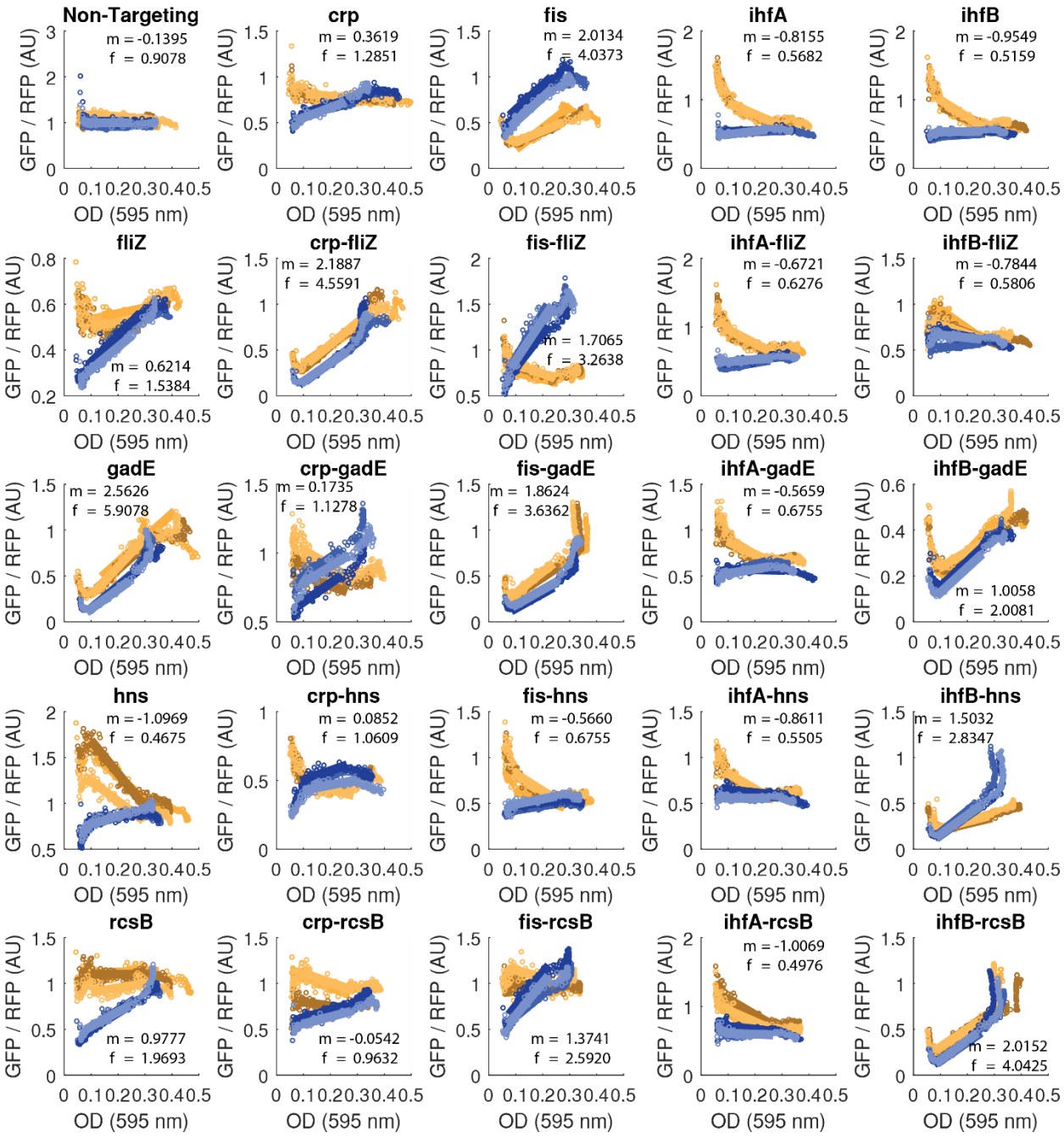



*Figure 17: Competition assays between LC-E24 :: dcas9 2tetO HK022 attB pCRRNA :: EGFP and LC-E24 :: dcas9 2tetO HK022 attB pCRRNA :: mCherry in M9 media with 0.4% Lactose. Strains harboring pCRRNA :: EGFP carried on of 24 targeting spacers or a non-targeting control (indicated at the top of the plots). The strain harboring pCRRNA :: mCherry contained a non-targeting control spacer and used as a reference strain. All EGFP strains were grown in competition with the reference strain, at a starting ratio of 1:1 made from dilution of 1/100 from overnight cultures. The change in the ratio of GFP to RFP signal with the change in optical density (OD) indicates the relative change in proportion of each strain as cultures grew. A slope was fitted to the data between the 25-75% of OD to avoid changes in the lag phase (which may be instrument noise) and stationary phase. A slope (m) of 0 indicates no change in relative proportions of each strain. A fitness score (f) was determined for each strain by computing 2^m. Data represents 2 biological replicates with 3 technical replicates each.*

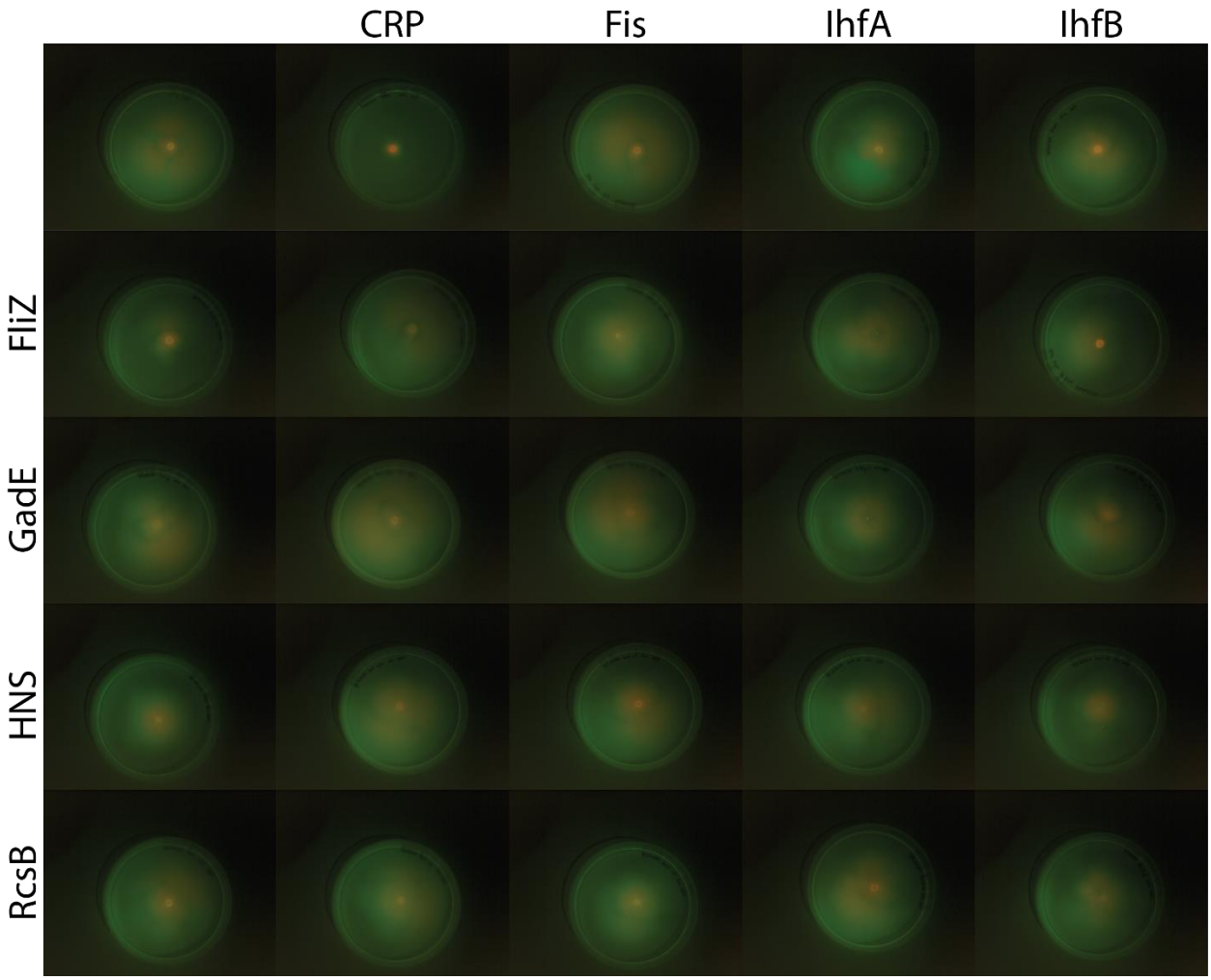



*Figure 18: Swimming assays of LC-E24 :: dcas9 2tetO HK022 attB pCRRNA :: EGFP and LC-E24 :: dcas9 2tetO HK022 attB pCRRNA :: mCherry in 0.3% LB soft agar. pCRRNA :: EGFP vectors contained one of 24 targeting spacers (corresponding to rows and columns) or a non-targeting control (uppermost left corner). The strain containing pCRRNA :: mCherry always contained a non-targeting vector for use as a reference strain. A bright center dot indicates cells that were unable to penetrate the agar due to a lack of any swimming ability.*

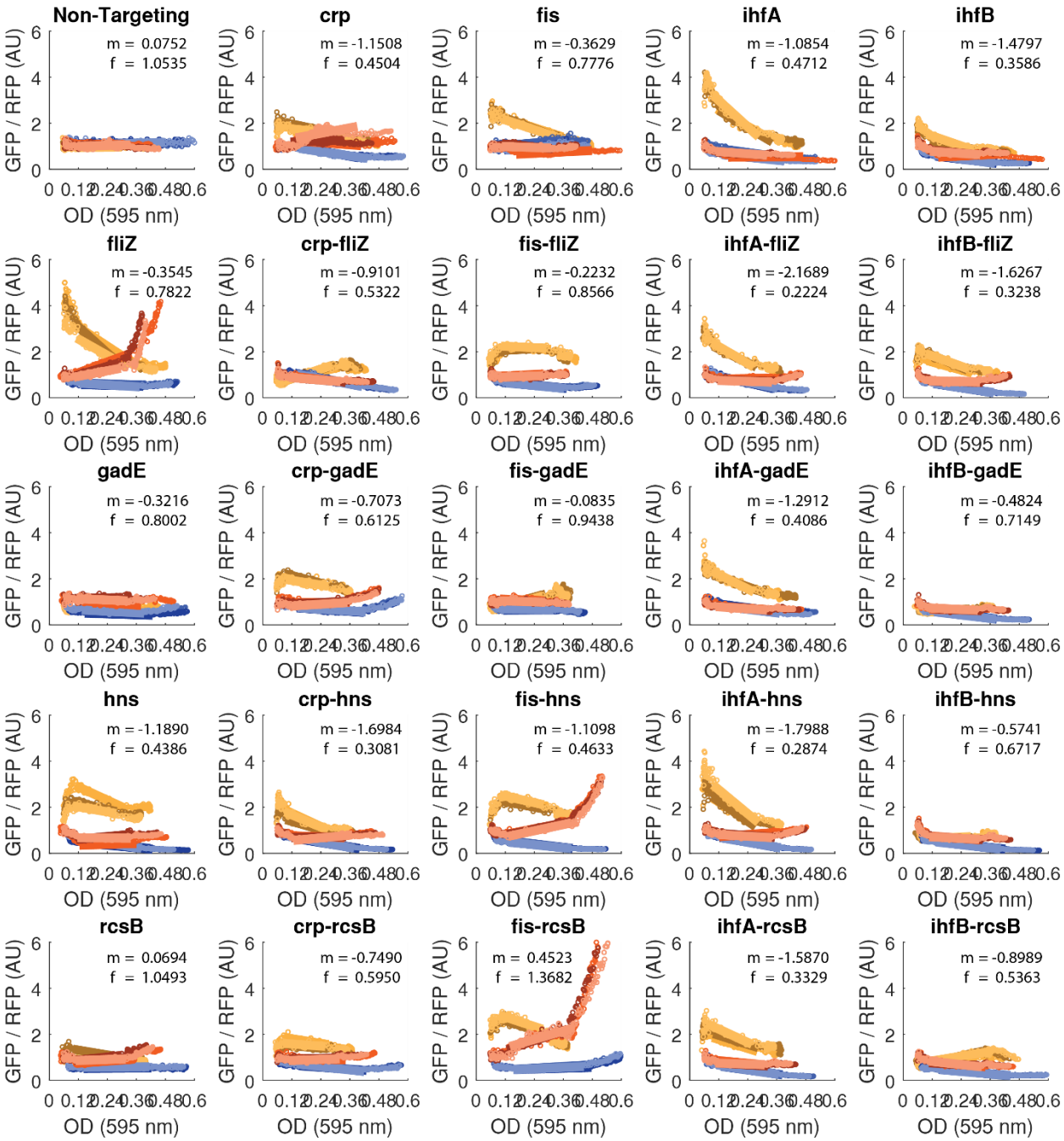



*Figure 19: Competition assays between MG1655 pdCas9 pCRRNA :: EGFP and MG1655 pdCas9 pCRRNA :: mCherry in M63 media with 0.4% glucose. Strains harboring pCRRNA :: EGFP carried on of 24 targeting spacers or a non-targeting control (indicated at the top of the plots). The strain harboring pCRRNA :: mCherry contained a non-targeting control spacer and used as a reference strain. All EGFP strains were grown in competition with the reference strain, at a starting ratio of 1:1 made from dilution of 1/100 from overnight cultures. The change in the ratio of GFP to RFP signal with the change in optical density (OD) indicates the relative change in proportion of each strain as cultures grew. A slope was fitted to the data between the 25-75% of OD to avoid changes in the lag phase (which may be instrument noise) and stationary phase. A slope (m) of 0 indicates no change in relative proportions of each strain. A fitness score (f) was determined for each strain by computing 2^m. Data represents 3 biological replicates with 3 technical replicates each.*

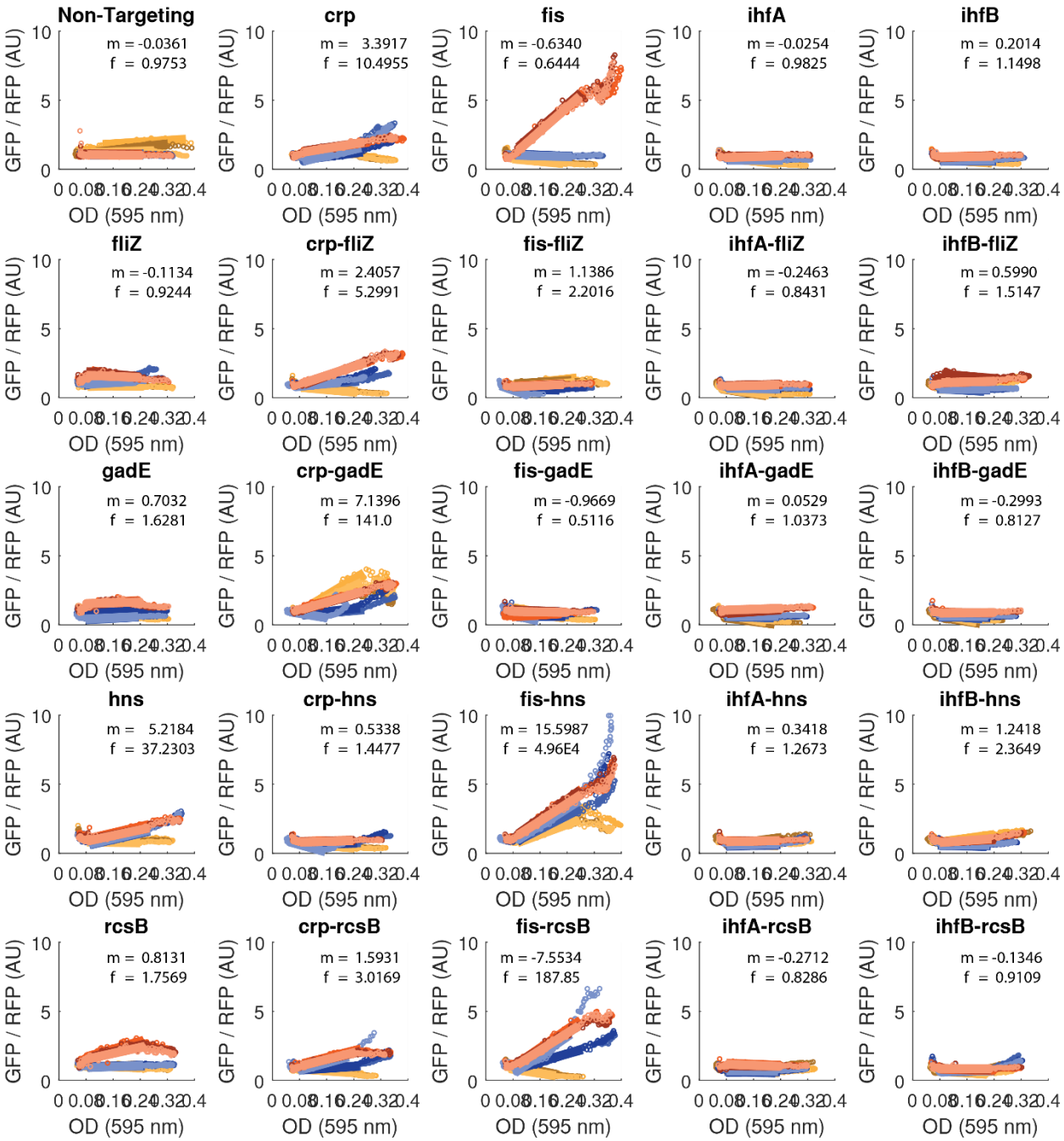



*Figure 20: Competition assays between MG1655 pdCas9 pCRRNA :: EGFP and MG1655 pdCas9 pCRRNA :: mCherry in M63 media with 0.4% lactose. Strains harboring pCRRNA :: EGFP carried on of 24 targeting spacers or a non-targeting control (indicated at the top of the plots). The strain harboring pCRRNA :: mCherry contained a non-targeting control spacer and used as a reference strain. All EGFP strains were grown in competition with the reference strain, at a starting ratio of 1:1 made from dilution of 1/100 from overnight cultures. The change in the ratio of GFP to RFP signal with the change in optical density (OD) indicates the relative change in proportion of each strain as cultures grew. A slope was fitted to the data between the 25-75% of OD to avoid changes in the lag phase (which may be instrument noise) and stationary phase. A slope (m) of 0 indicates no change in relative proportions of each strain. A fitness score (f) was determined for each strain by computing 2^m. Data represents 3 biological replicates with 3 technical replicates each.*

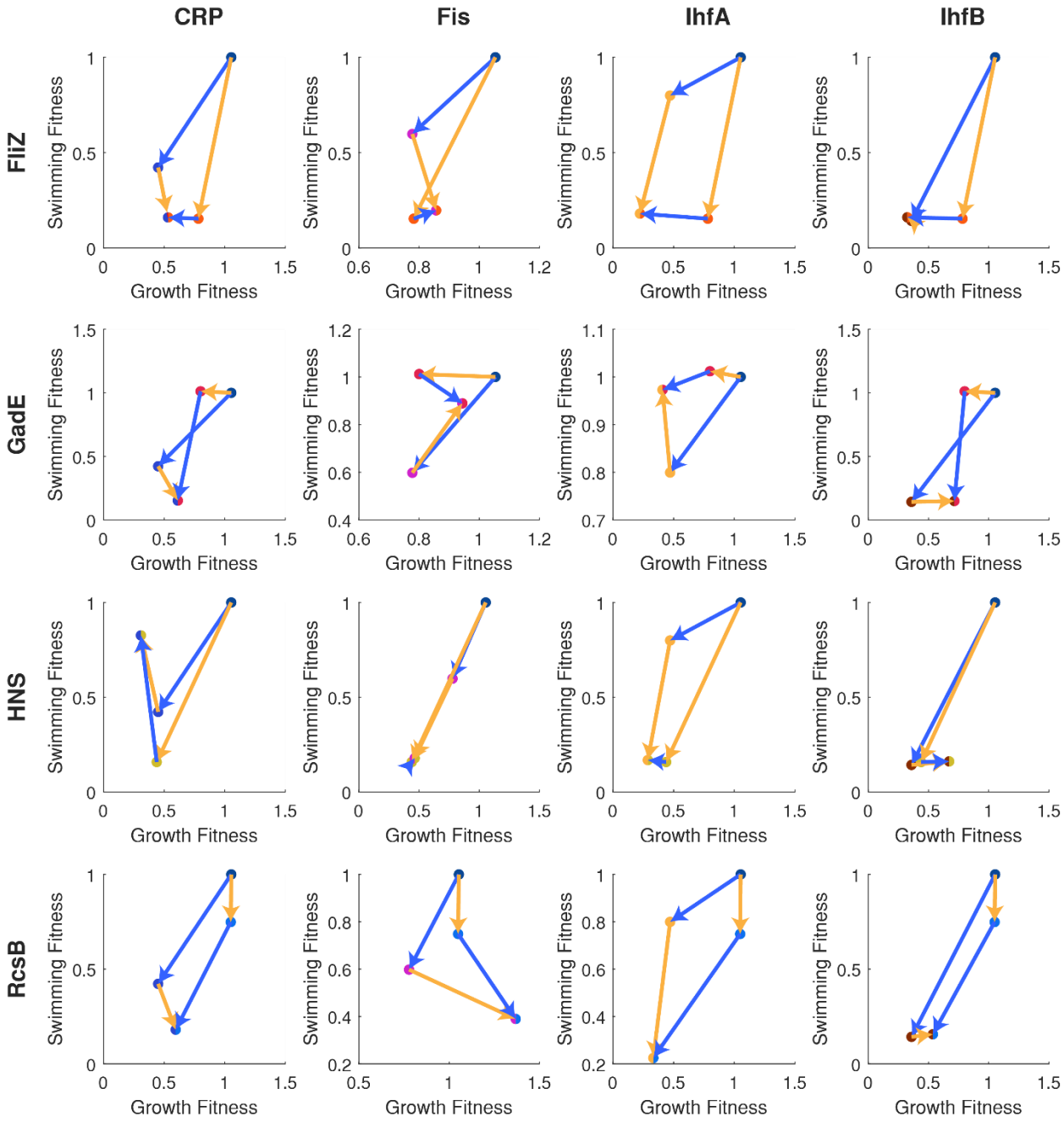



*Figure 21: Fitness trajectories for MG1655 pdCas9 pCRRNA :: EGFP strains in M63 media with 0.4% glucose. Fitness measurements for each crRNA spacer were normalized to the reference strain (non-targeting spacer) such that the reference strain had a fitness of 1. Single and double knock-down mutants were then plotted for each combination of strains. Blue arrows indicate the fitness trajectory with the addition of the upstream (columns) gene while yellow arrows indicate the trajectory with the downstream (rows) genes. Trajectories with opposite directions in one or two axis represent sign epistasis for the respective fitness.*

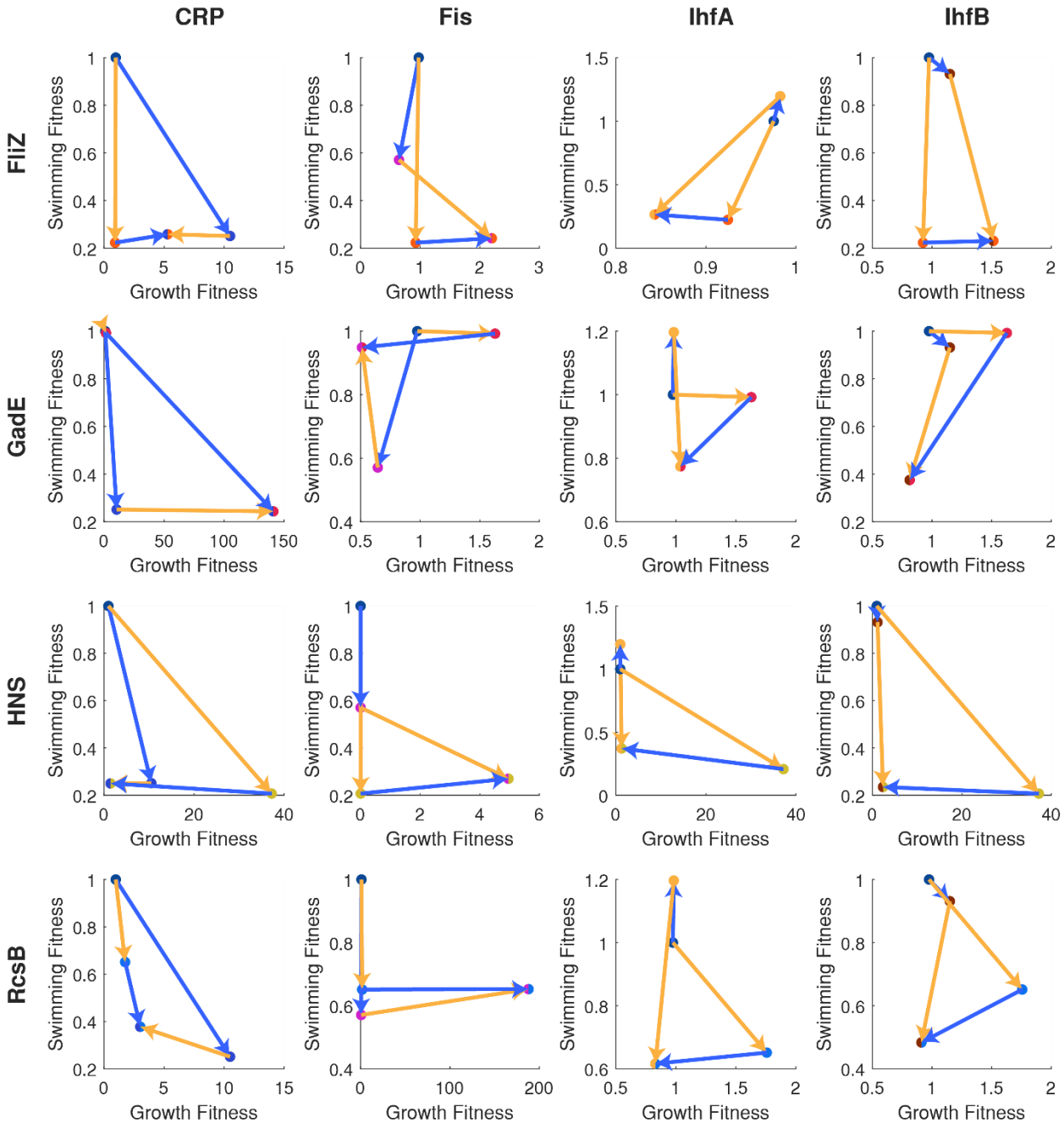



*Figure 22: Fitness trajectories for MG1655 pdCas9 pCRRNA :: EGFP strains in M63 media with 0.4% lactose. Fitness measurements for each crRNA spacer were normalized to the reference strain (non-targeting spacer) such that the reference strain had a fitness of 1. Single and double knock-down mutants were then plotted for each combination of strains. Blue arrows indicate the fitness trajectory with the addition of the upstream (columns) gene while yellow arrows indicate the trajectory with the downstream (rows) genes. Trajectories with opposite directions in one or two axis represent sign epistasis for the respective fitness.*

# 3 Multiplexed Knockdowns of Global Transcription Regulators

A small number of transcription factors drive expression for most of the genes in *E. coli.* The transcriptional regulatory network follows a power-law distribution, where a few nodes have many connections and most nodes only have a few connections [37]. Additionally, these transcription factors group into functional modules [40]. How can we reconcile this complex regulatory structure with the observations of a two state response [42] [53]? Functional regulatory modules do not appear to be highly conserved, nor can transcriptional units be unambiguously linked to their direct regulators [40]. This brings us back to the question, is there a relationship between the connections in transcriptional regulatory network and the response of the system to perturbations?

To understand how global regulators work in concert to alter gene expression at a genome wide scale, we perturb all possible combinations of 5 global transcriptional regulators *arcA, crp, fis, fnr,* and *hns*. We first record fitness measurements for all of the resulting 32 strains in 3 different growth media. Two important observations are that i) the fitness trajectories are not monotonous. That is to say that changes in fitness occur in both directions as we increase the number of perturbations within a strain. And ii) the fitness profiles of the strains are media dependant. We then calculated the epistasis between these five global regulators and find that there are higher order epistasis in both fitness measurements we use, and the order of epistasis (that is the number of epistasis terms required to explain the data) is consistent across growth media. These high-order interactions are significant as they shape the accessible evolutionary paths of the genes [56].

We recorded the transcriptional profile of all of our strains in both the exponential and stationary phases of growth. We performed principle component analysis to analyse the dimensionality of the genetic response. These principle components attempt to explain the variance of the expression data by reducing correlated genes into a new 'component' to represent that variance. This effectively reduces a highly correlated, high dimensional data set such as RNA-sequencing data into a smaller number of uncorrelated 'principle components'. Here we can compare the dimensionality of the inputs to the system (the number of perturbed genes) to the dimensionality of the output of the system (the number of principle components which represent most of the data) similar to an IN-OUT system in engineering. Consistent with previous results [42] [53], we find that the growth phase is the largest contributor to the variance in gene expression. We also find that each individual perturbation is strongly associated with one of the first 7 principle components. However these 7 components only account for 59% of the variance in the data, and we need 22 components to account for over 75% of the variance.

We also performed cluster analysis on the transcriptional data to identify common genetic programs, and attempt to dissect the logic which governs which sets of perturbations lead to these programs. Finally, we discuss a logical model [59] for the transcriptional network, here we also see that perturbations increase the dimensionality of the system, and how logical data from our transcriptional data could be fed into this model.

## 3.1 CRISPR-Cas9 Knockdowns

*With **Marie Baumont***

***Single CRISPR spacers target E. coli genome while non-targeting control spacer does not***

We designed, cloned and tested CRISPR targeting sequences for the global transcriptional regulators *arcA, crp, fis, fnr,* and *hns.* CRISPR targeting sequences are designed by first identifying all PAM (NGG) sequences in the promoter region of each gene for each strand of DNA. The 30 nt flanking the PAM sequence is then labelled as a potential CRISPR Spacer. The Spacers are then checked for potential off targeting by searching all other instances of the last 10 nt on the 3' end of the spacer within the genome of MG1655. Only the 3' end of the spacer is considered as mutations are tolerated much more on the 5' end of the spacer than the 3' end [71] [80] [77]. Each matching 10 nt sequences is then checked for a flanking PAM sequence and if one is found, that spacer is labelled with potential off target effects. CRISPR targeting spacers with no predicted off target effects that are located on the bottom template strand near the σ70 start site are preferentially selected for cloning into pCRISPR.

Single CRISPR Knockdown Vectors were transformed into strains MG1655 :: pTet Cas9 and MG1655 :: pTet dCas9. Growth curves were performed on each strain with induction by aTc to determine if CRISPR targeting spacers were functional as functional Spacers should be lethal in MG1655:: pTet Cas9 but not dCas9 (Figure 23). Critically, no difference in growth rate was observed between Cas9 and dCas9 with the non-targeting control strain, while differences were observed in all the targeting strains. Next qRT-PCRs were performed on MG1655 :: dcas9 strains with pCRISPR vectors for each single knock down to determine the level on inhibition of each gene (Figure 24). We reliably see a shift in gene expression for each spacer, even those that did not have a clear phenotype in the growth curve assays.

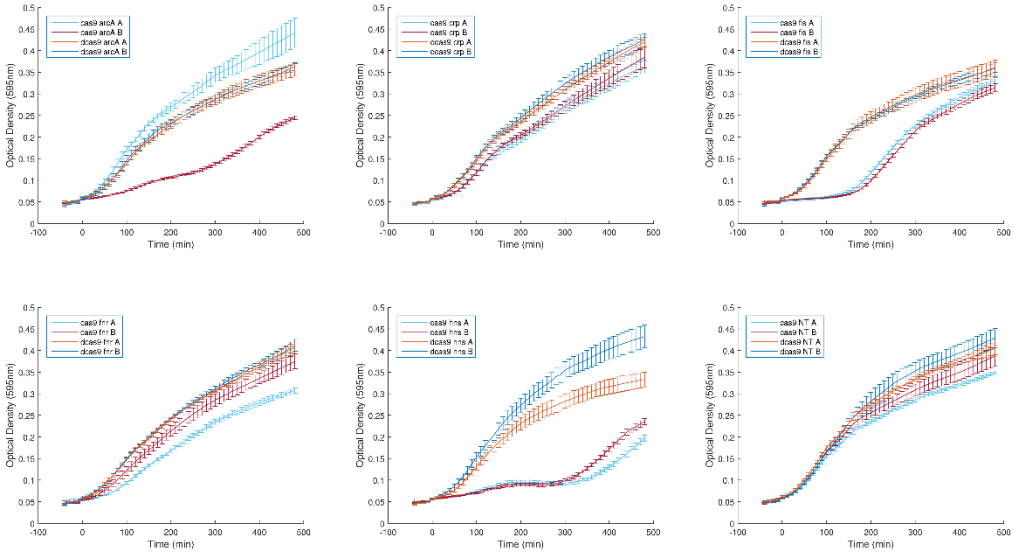



*Figure 23: Growth Curves of Cas9 and dCas9 strains containing crRNA with targeting sequences for arcA, crp, fis, fnr, and hns. The double strand DNA breaks caused by Cas9 result in a strong fitness decrease of each of the targeting spacers. The non-targeting control does not have any known targets in E. coli MG1655.*

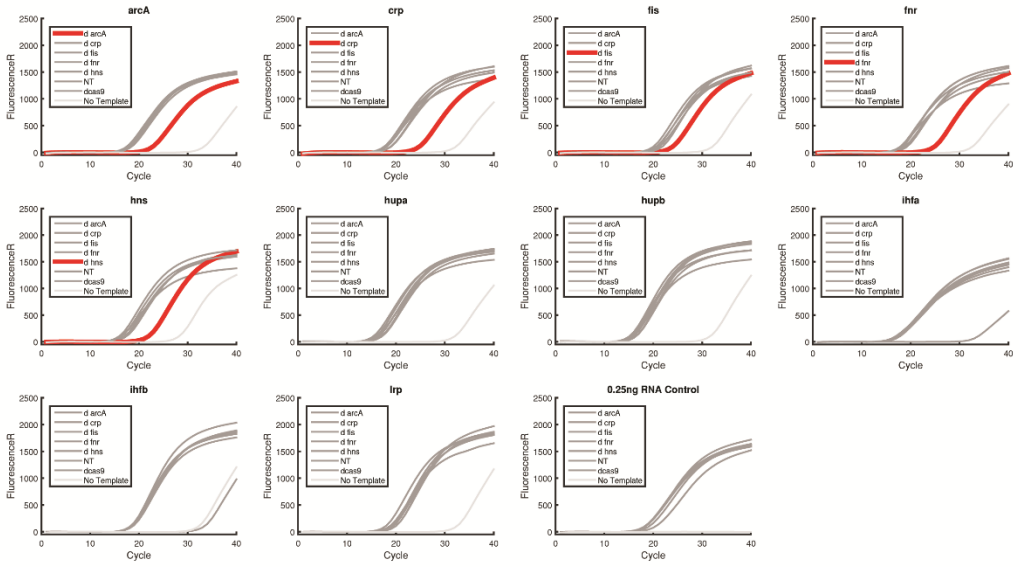



*Figure 24: qPCR on 7 strains of E. coli expressing dCas9. Five strains carry a targeting crRNA for either arcA, crp, fis, fnr, or hns. A sixth strain carries a non-targeting spacer, and a seventh has no crRNA. The cDNA generated for the global regulators was amplified after RT-PCR by qPCR. The highlighted red curve indicates the strain which carries the perturbation for that gene.*

### *Quantification of CRISPR perturbations on gene expression with qRT-PCR*

We initially used a panel of genes as reporters for quantitative reverse transcription polymerase chain reaction (qRT-PCR). This allows us to quantify the relative number of transcripts in each strain. The gene panel consisted of genes within the transcriptional regulatory network of *E. coli* which were directly regulated by our targeted genes. This list consisted of arcA, crp, fis, fnr, hns, hupa, hupb, ihfA, ihfB, lrp, marRA, gadX, ompR, oxyR, fur, gadE, csgD, and flhDC. We extracted RNA from strains containing each of our perturbations as well as 3 control strains (no crRNA and non-targeting crRNA as negative controls, and a genetic knock-out of hns as a positive control). RNA was extracted at multiple growth stages as determined by the optical density. Once RNA was extracted, we performed our qRT-PCR reactions to determine the number of transcripts for each gene in each sample. We then performed dimensionality reduction with the qPCR data using principle component analysis (PCA). This allowed us to reduce the dimensionality of our data from 18 (the number of genes measured). This allowed us to apply the same techniques that were used a microarray dataset [132] and in our larger screens with a smaller set of data.

When we colored our data by the optical density of the sample in the first three principle components, we found that the first component corresponded to the optical density, with samples in exponential phase had positive principle component scores for the first component, while samples in early stationary phase had negative scores (Figure 25). This was consistent with the finding with microarray data that the first principle component was strongly associated with growth stage [42] [132]. We then colored the samples by their perturbations caused by the crRNA (Figure 26). We found that samples with

the same perturbations were close to each other in the first three principle components. Additionally, the two negative controls were both close to each other as well, indicating that the non-targeting crRNA wasn't having an effect on gene expression. The samples with the strongest negative scores for the second principle component were all perturbed for hns. This was slightly surprising as all of the other perturbations are higher in the regulatory network hierarchy than hns, and thus regulate the expression of more of the genes in our panel, however hns is known to have a strong effect on over-all fitness effect as it represses many stress response genes [133]. Additionally, more of our samples had hns perturbed than the other crRNA perturbations given the positive control for hns. In the third component, we saw that strains with fnr perturbed had strong scores compared to the rest of the samples. These results indicate that at least two of the perturbations seem to be acting independently in our qRT-PCR data set.

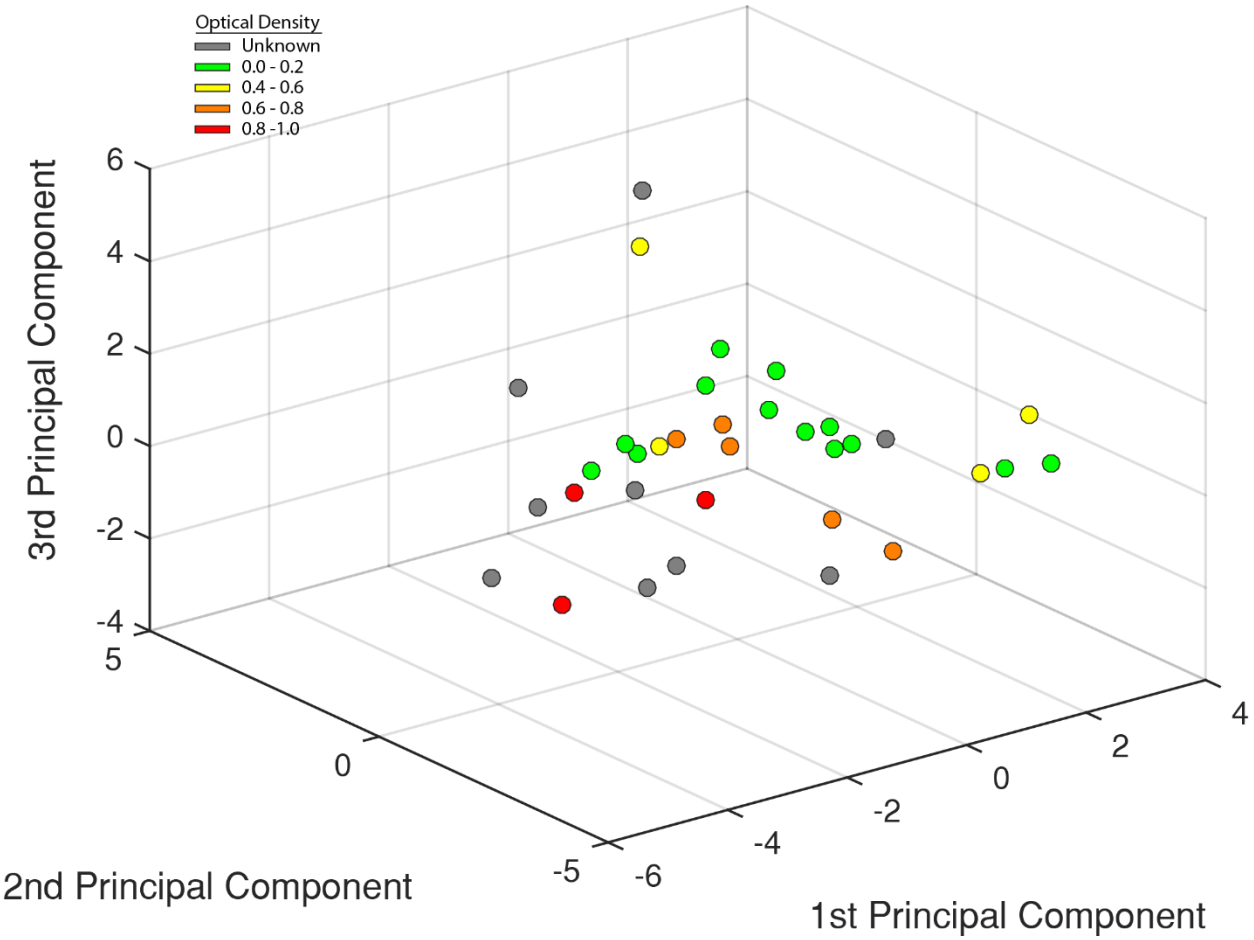



*Figure 25: Principle Components of qRT-PCR data from perturbed cells, Colored by the Optical Density (OD) of the Cell Culture. Transcription Factors directly regulated by perturbed genes were quantified by qRT-PCR for cells perturbed with CRISPR-dCas9 at various ODs.*

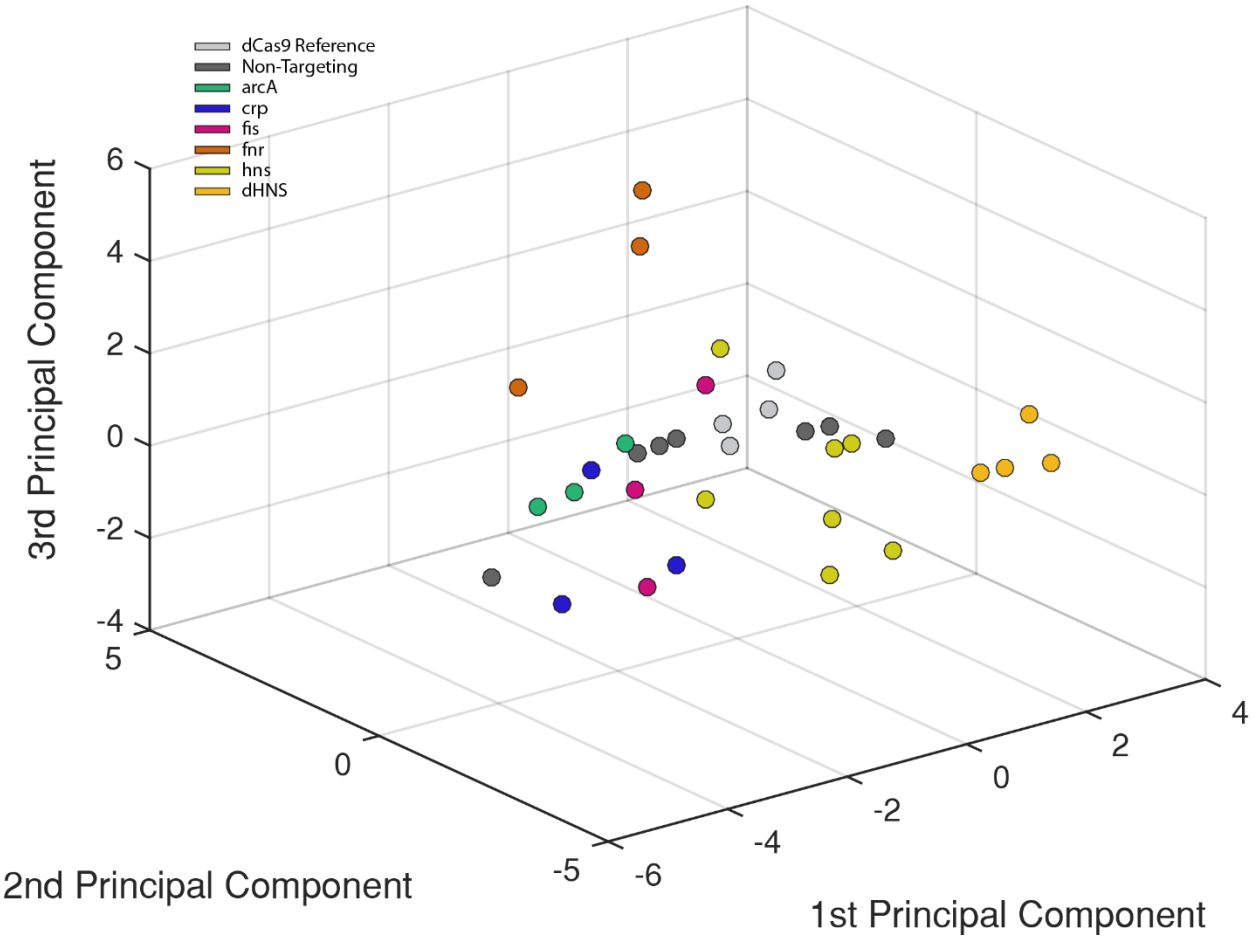



*Figure 26: Principle Components of qRT-PCR data from perturbed cells, Colored by the perturbation of the Cell Culture. Transcription Factors directly regulated by perturbed genes were quantified by qRT-PCR for cells perturbed with CRISPR-dCas9 at various ODs. A non-targeting crRNA strain and a dCas9 strain without a crRNA were taken as negative controls, while a strain with HNS knocked out from the KIEO collection was taken as a positive control (dHNS).*

***CRISPR spacers repression is optimal when targeting the coding strand at the transcriptional start site***

Initially, we chose targeting spacers close to the promoters for all genes except fis, which was targeting at the beginning of the open reading frame (ORF). This was due to the presence of dusB between fis and its transcriptional start site. However, we noted lower fis repression than with other targeting spacers. Part of the cause of this is that fis expression lowers rapidly after early stationary phase but determine if there were other targeting site that could result in a more consistent repression, we designed four targeting spacers for each gene, two which annealed to the coding strand and two which annealed to the template strand. Two of these targeted the promoter regions and two targeted the beginning of the open reading frame. We then extracted RNA from strains expressing these crRNA and performed qRT-PCR to quantify the strength of the gene repression (Figure 27). Consistent with published results, we found that crRNA complementary to the coding strand repressed expression better than crRNA that targeted the template strand [80] [77] [16]. In genes with the promoter directly in front of the ORF, there was little difference between the spacers complementary to the coding strand, regardless of whether they targeted the promoter or the ORF. However in the case of fis, we found that repression was significantly worse when targeting the ORF. Even though it may confound our results due to the repression of dusB,

we decided to change our targeting crRNA for fis to target the promoter region to have a stronger and more reliable repression of the fis gene. This is likely to have little impact on the cell phenotype, as DusB is redundant with DusA and DusC. Additionally, DusB is synthesized at very low levels compared to Fis, despite being located within the same operon, with less than 0.5% of the protein levels of Fis [44].

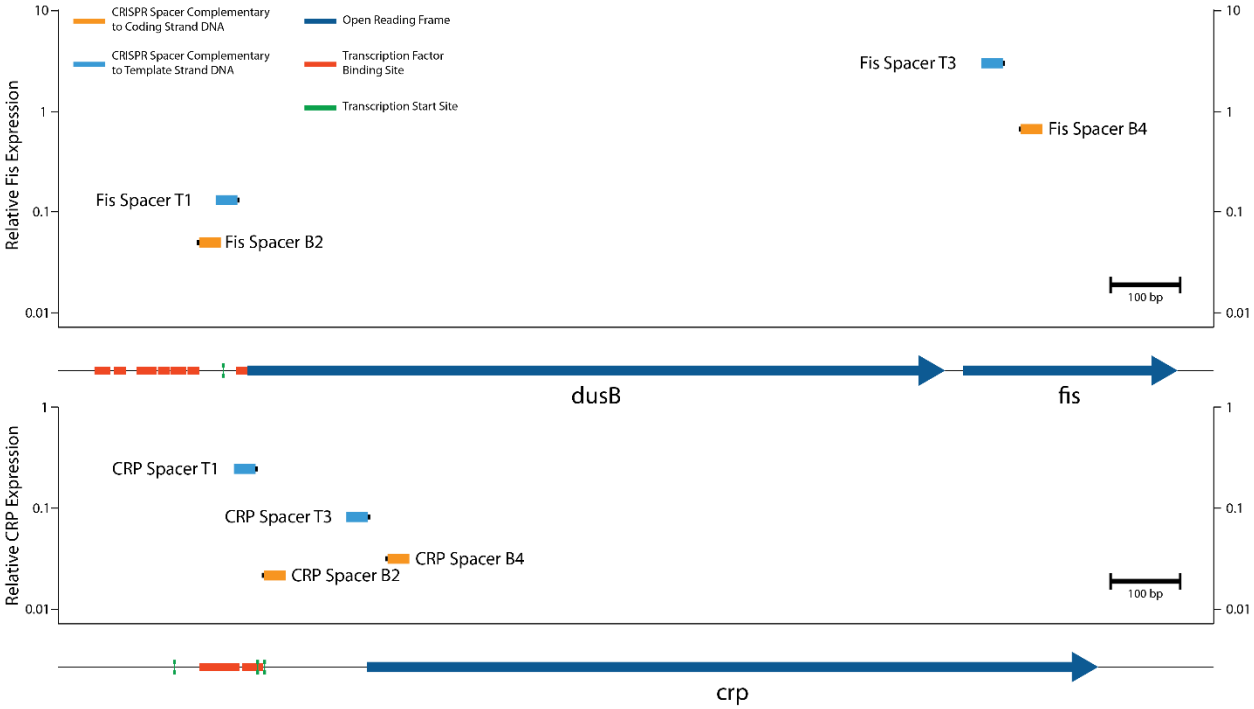



*Figure 27: Multiple spacers were designed for each gene on that annealed to either the coding strand (orange) or the template strand (Blue). The black dot on the side of a spacer indicates the NGG PAM sequence. The position of each spacer was mapped to the genetic sequence with open reading frames annotated as dark blue arrows, transcriptional start sites in green, and transcription factor binding sites in red. Relative gene expression was determined by qRT-PCR by comparing the Cq of perturbed cell lines with the Cq of a non-perturbed control.*

## 3.2 CKDL Vector for Multiplexed Knockdowns

***Multiplexing CRISPR perturbations with pCKDL vector***

Our approach to perturb the genetic regulatory network of *E. coli* relies on the programmable nature of the CRISPR-Cas9 system. Here, a catalytically deactivated Cas9 enzyme (dcas9) is used as a DNA binding protein, which is able to act as a transcriptional repressor when it binds to the promoter region of a gene. The binding location of dCas9 is determined by a small guide RNA called crRNA which naturally exists in an array of many targeting spacers. We take advantage of this programmable nature by building all possible combinations of targeting RNA so that we can fully explore the interactions between nodes within a network (Figure 28). The crRNA is naturally able to target multiple sequences to provide protection from a variety of phages with some natural CRISPR locus containing hundreds of spacers [134]. We can take advantage of this feature by putting multiple targeting spacers in each CRISPR locus to multiplex perturbations of a genetic network. We can consider the perturbations as a Boolean function, with zero and one being non-perturbed and perturbed respectively. We then create a CRISPR array with a number of spacers equal to the number of genes we wish to investigate, with each corresponding to a specific gene.

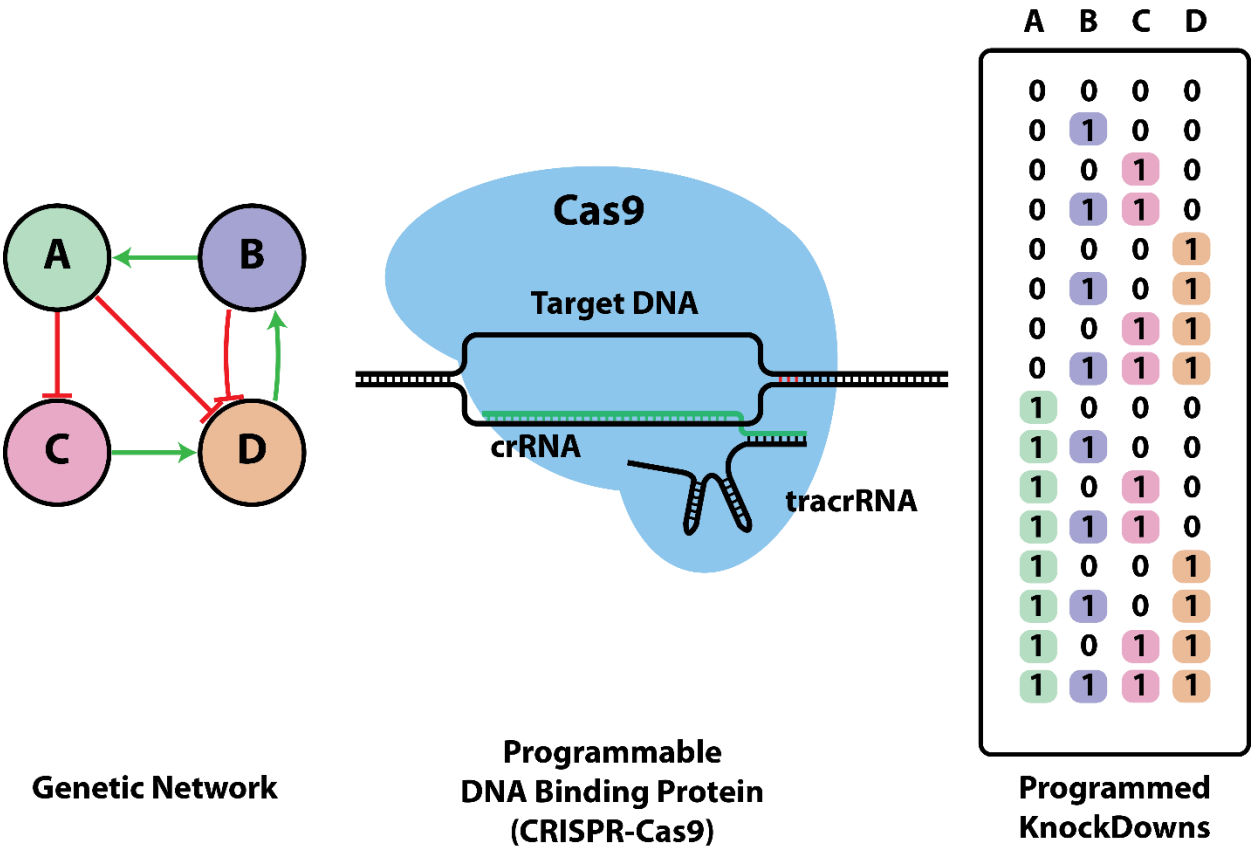



*Figure 28: Combinatorial Knockdown strategy using CRISPR-Cas9. To explore a genetic network, a program for targeting all possible combinations of genes can be created using CRISPR-Cas9. Each gene is represented in the CRISPR array as zeros and ones, as either a non-targeting spacer or a targeting spacer respectively.*

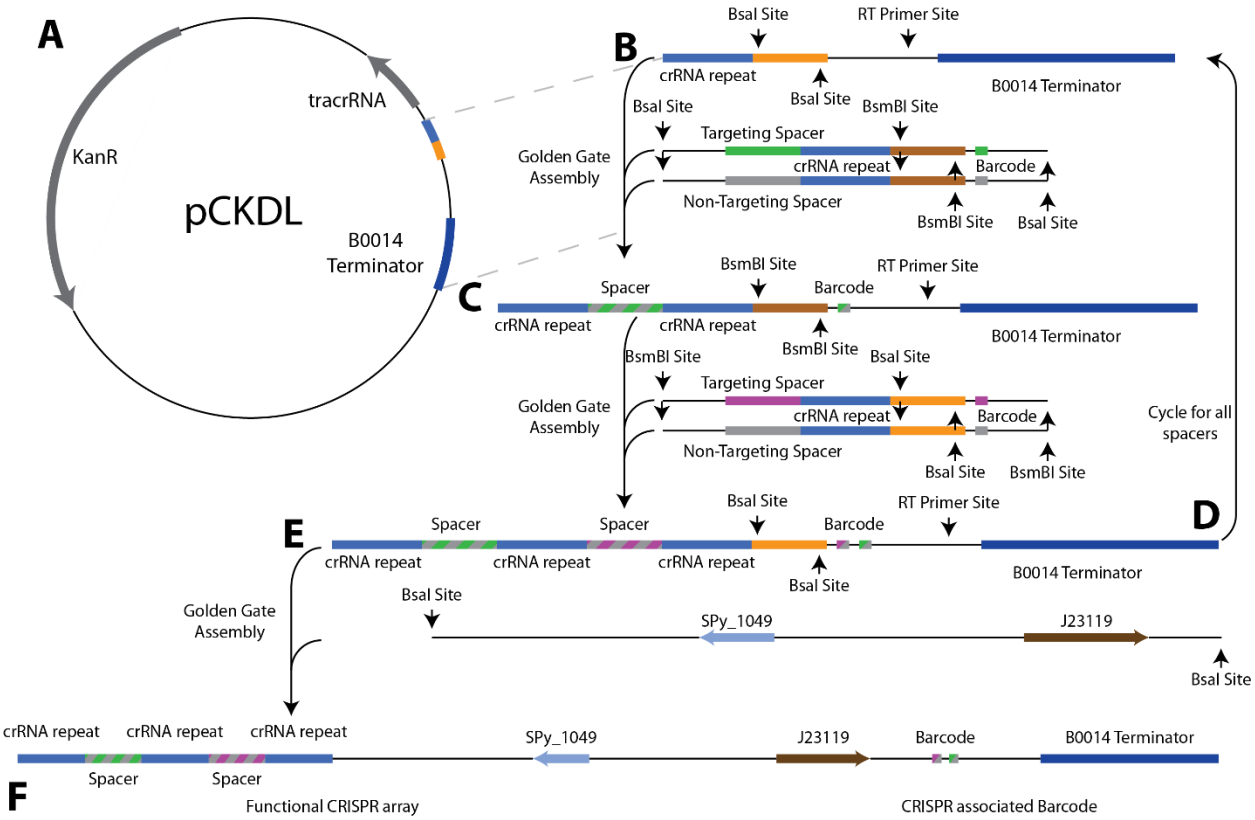



*Figure 29: Cloning with pCKDL system. The pCKDL vector is derived from pCRRNAcos and contains a tracrRNA, part of a crRNA, a cloning site, B0014 Terminator and the Kanamycin resistance marker (**A**). The pCKDL vector contains 2 BsaI cut sites next to a crRNA repeat sequence with no leader sequence or promoter(**B**). The BsaI sites allow for golden gate assembly with dsDNA fragments containing a DNA barcode, a BsmBI cloning site, a crRNA repeat, and either a spacer targeting a gene of interest or a non-targeting spacer. Golden gate assembly with these dsDNA fragments extends the crRNA array in the vector with either spacer and changes the cloning site to BsmBI(**C**). This allows golden gate assembly to be repeated with another set of dsDNA fragments, again with either a targeting or non-targeting spacer, but with a BsaI cloning site within them. This returns the vector to the initial cloning configuration, allowing the process to be cycled for the desired number of total targets (**D**). When the desired number of targets have been inserted, the crRNA lead sequence, as well an PCR handle or a promoter to drive expression of the barcode (for scRNAseq) can be inserted with golden gate assembly as well(**E**). The completed library of vectors (**F**) contain all possible combinations of all spacers, without any redundancies.*

We use a modified pCRRNA [135] vector for perturbing multiple genes in *E. coli* (called pCKDL). This vector has been modified to remove half of the crRNA and replace it with a terminator and RT binding site. The CRISPR array is then built up through subsequent rounds of cloning with gBlocks inserted into a cloning site within the CRISPR array. Each gBlock contains a CRISPR Spacer, a Repeat, a new cloning site, and a DNA barcode. In each round, there will be a targeting spacer for a specific gene, or a non-targeting spacer added to each vector. This allows us to build a library of CRISPR knock-down mutants at a size of 2^n, where n is the number of targeted genes. As such, the library avoids redundant mutants that typically occur when all possible genes are cloned at each cloning step [136]. The cloning site with the vector alternates between a BsaI site and an Esp3I site with each round, allowing for a cycling golden-gate-like assembly method (Figure 29). We have constructed a library of 32, with 5 perturbation targets: ArcA, CRP, Fis, FNR, and HNS.

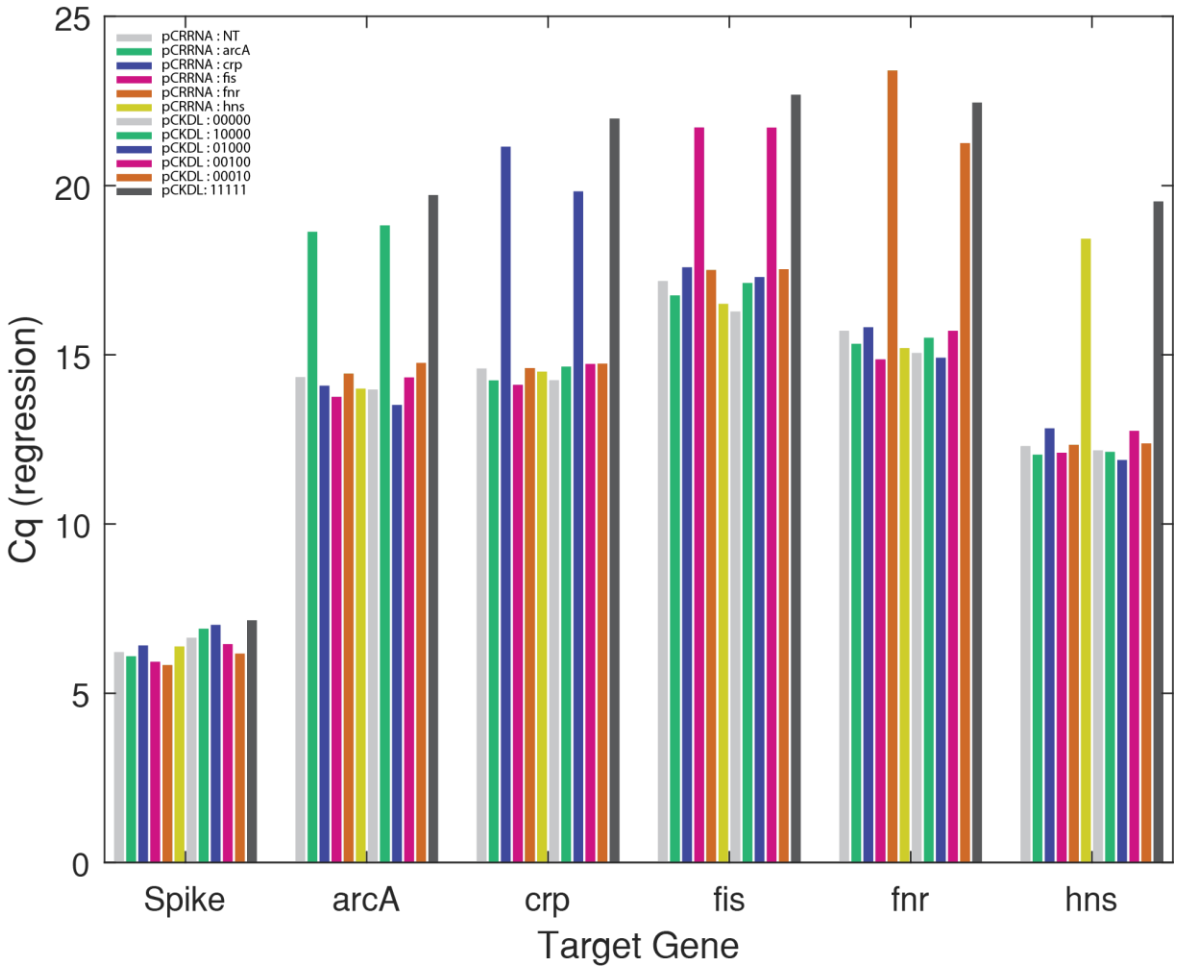



*Figure 30: Comparison of Gene Repression between pCRRNA and pCKDL with a single targeting spacer or multiple targeting spacers. E. coli strain LC-E24 harboring either pCRRNA with a single spacer or pCKDL with 5 spacers, colors correspond to targets of the spacers (green: arcA, blue: crp, pink: fis, orange: fnr, yellow: hns). pCKDL vectors contain either a single targeting spacer and 4 non-targeting spacers, or a targeting spacer for all 5 targets (dark grey). Vectors with only non-targeting spacers used as controls (light grey). qRT-PCR was preformed on RNA extracted from strains during exponential growth phase and an RNA spike was added to detect differences in RT efficiency.*

***Assembly of pCKDL vectors and validation they function as efficiently as single targeting vectors***

We began by inserting a spacer targeting ArcA or a non-targeting crRNA spacer into pCKDL using GoldenGate Assembly and the BsaI enzyme (Figure 16). This resulted in two nearly identical plasmids pCKDL 0 and pCKDL 1 where the only difference is in the content of the single crRNA spacer and the 5nt DNA barcode. It is useful to visualize a library as a binary number with each digit representing a spacer in the crRNA array. If the crRNA contains a spacer that targets a gene, the digit corresponding to that spacer is a 1, and if it does not target a gene the digit is a 0. In this way all the binary numbers from 0 to 2N can represent all the plasmids we construct with pCKDL. With each of the plasmids constructed from round one, we then inserted a spacer targeting CRP or a different non-targeting spacer; using GoldenGate Assembly and the BsmBI enzyme. This results in 4 unique plasmids; pCKDL 00, 01, 10, and 11. Importantly, while all of these plasmids have unique combinations of spacers and DNA barcodes, they all have identical BsaI cloning sites. At this point round 1 and round 2 cloning steps can be alternated until a desired library size is reached, each time simply replacing the targeting spacer with a new gene. For our library of 32

constructs we kept each plasmid separated in each step so that bulk experiments could be performed. However for high-throughput experiments, it is possible to pool the resulting plasmids after each round to prevent the number of parallel cloning reactions from becoming overwhelming. Sequencing the DNA barcode can then be used to determine what combination of spacers is contained in each vector.

When the CRISPR array has been constructed, the fragment of pCRRNA that was removed to make pCKDL is reintroduced, and a promoter can be added with it to express the DNA barcode as RNA for detection during reverse transcription. When we finished construction of our 32 plasmids, we performed qRT-PCR on a subset of them to ensure that the modifications we made to pCRRNA when making pCKDL hadn't interfered with expression of the crRNA or the tracrRNA, and that when using multiple spacers in a single crRNA, we were not titrating out the dcas9 protein or the tracrRNA and reducing the strength of our repression. As such we compared pcrRNA with spacers for ArcaA, CRP, Fis, FNR, and HNS to pCKDL 10000, 01000, 00100, 00010, and 00001 respectively, along with pcrRNA with a non-targeting spacer and pCKDL 00000, and 11111 (Figure 30). We found expression of targeted genes to be comparable between pCRRNA and pCKDL with no loss of repression. We sequenced our entire 32 plasmid library to ensure there were no errors or mutations, finding only that pCKDL 11110 had spontaneously lost its non-targeting spacer. We also noted that expression of the DNA barcode as a small RNA had a noticeable fitness cost, and such for our bulk experiments with fitness and RNA sequencing we created our final pCKDL without a promoter for the barcode, although this will be required for droplet based single cell RNA sequencing.

## 3.3 Fitness and Epistasis of pCKDL Perturbations

***Growth curves of pCKDL show patterns of growth effects corresponding to perturbations***

Once we finished construction of our pCKDL strains, we determined the fitness measurement of each strain by 20H growth curves in 96 microtitre plates. Strains were grown in either LB or M9 media, with M9 media supplemented with either 0.4% Glucose or Lactose as a carbon source. The Cas9 enzyme was induced at the beginning of each growth curve. We chose two fitness metrics to quantify differences in the growth curves between strains (Figure 31). First, a growth rate fitness, is the slope of growth curve during exponential growth. The location where the slope is measured corresponds to the maximum

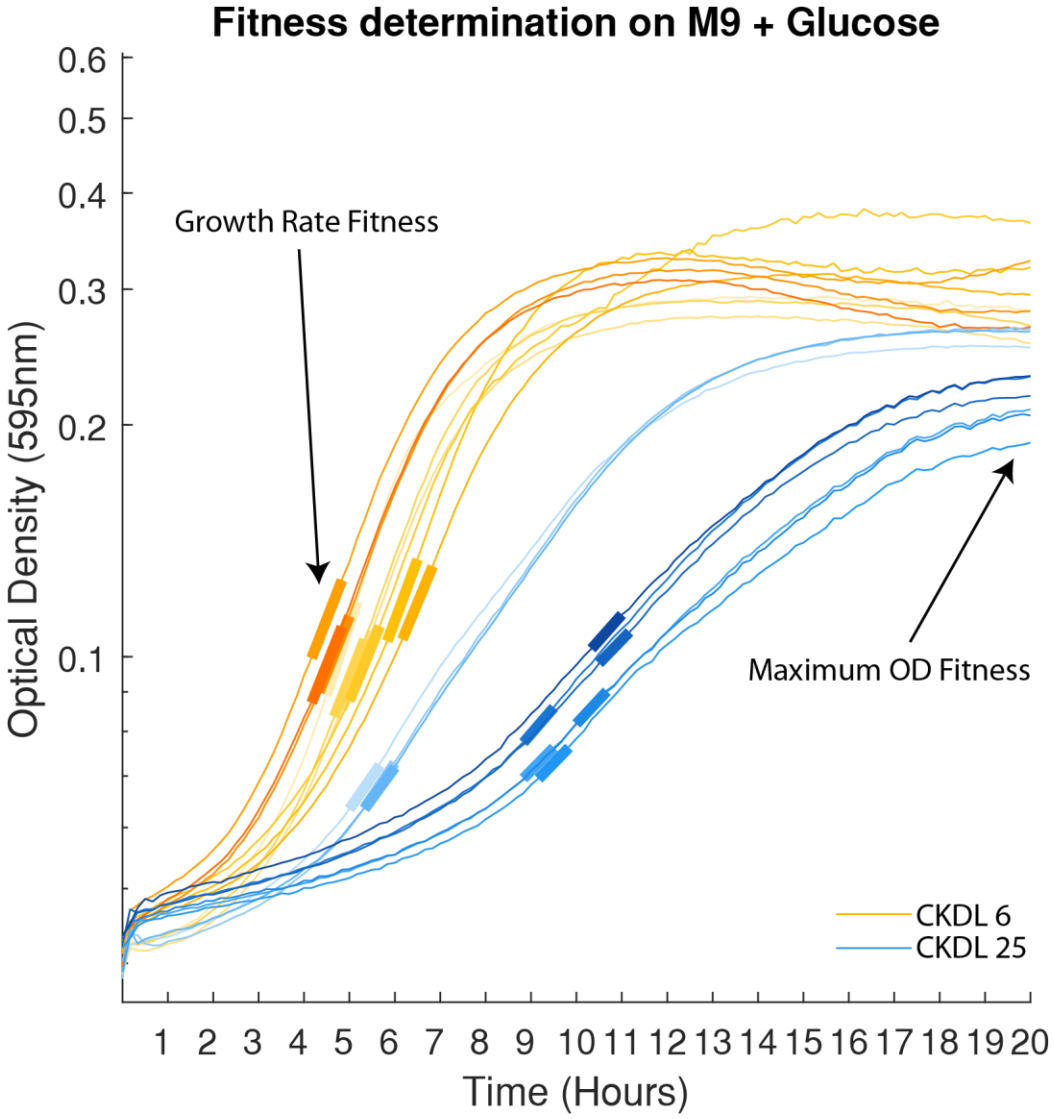



*Figure 31: Fitness Determination from Growth Curves. Growth curves for two strains (phenotype extremes) of LC-E24 bearing pCKDL 6 (orange) and 25 (blue) grown in 200 µL of M9 Media supplemented with 0.4% Glucose. Fitness is determined by taking the derivative of the log2 of the growth curves to find the exponential growth phase. A linear fit is made to the growth curve ± 20 minutes of the maximum derivative (Bold section of curve). The slope of this linear fit is taken as the Growth Rate Fitness. The second phenotype measured is the maximum OD reached after 20 hours. Each strain has these two fitness measurements in three media: LB, M9 + Glucose, and M9 + Lactose. There are 3 biological replicates (taken from independent single colonies), and three technical replicates for each biological replicate.*

derivative of the growth curve. Second, a maximum Optic Density (OD) fitness, refers to the maximum $OD_{595nm}$ that is reached after 20 hours of growth. These fitness measurements were measured for all 32 strains, in 3 different media, with 3 biological replicates and 3 technical replicates for each biological replicate. All 32 strains were able to grow well in LB (Figure 49). Due to the difficulty in visualizing individual growth curves among 32 strains, we also represented the growth curves as parallel rows in a heatmap, with the intensity corresponding to the Optic Density (Figure 50). From this perspective there is a clear striped pattern in which the OD alternates up and down with each strain in stationary phase. This pattern corresponds to the absence or presence of the hns spacer in the pCKDL vector respectively. There also seems to be a pattern of every 8 strains where the time the OD shifts from an OD of 0.35 to 0.4 shifts from sooner to later corresponding to the presence or absence of the crp spacer in the pCKDL vector respectively.

***Fitness landscapes of pCDKL strains show that fitness effects are non-monotonous***

To explore this further we plotted the two fitness measurements for each strain (Figure 32). We colored each point by the spacers that were contained in the pCKDL for a given vector (with grey representing the pCKDL with no targeting spacers). Strains which vary by a single spacer in pCKDL are connected with an arrow, in the direction of the additional targeting spacer and colored by which spacer is added (green for arcA, blue for crp, pink for fis, orange for fnr, and yellow for hns). If there is a significant difference between the fitness of the two strains (as determined by Welch's t-test) the arrow is a solid line, and if not the arrow is dashed. Firstly, while the non-targeted reference strain pCKDL 0 has a high fitness for both growth rate and maximum OD, there are other strains with higher measurements for both fitness measurements. Some of these high performing strains include 4 targeting spacers. Additionally, the strain with all 5 targeting spacers, pCKDL 31, has a fitness very similar to the reference strain. The perturbations are non-monotonic, in that in general, the fitness initially decreases with the addition of targeting spacers, but then begins to increase the number of targeting spacers increases. Generally, most of the strains with low maximum OD fitness contain an HNS targeting spacer, and most of the strains with a low growth rate fitness contain an FNR targeting spacer.

To further visualize the effects of individual spacers on fitness, this data was decomposed into individual figures for each fitness measurement: Growth rate fitness (Figure 51) and Maximum OD fitness (Figure 52) where the fitness measurements were normalized to the references strain such that pCKDL 0 has a fitness of 1, and then plotted against the number of targeting spacers contained in the pCKDL vector. For both figures, connections between strains are colored and annotated as before. With the maximum growth rate, when FNR has been knocked down it significantly reduces the growth rate of the cell. This is also true for knock downs of fis and hns, though less common than fnr (3 and 4 incidents respectively compared to 6 for fnr). This contrasts with knock downs of crp and arcA, which increased growth rate in 3 cases each. There is only one pCKDL where the direction of the effect of spacers changed, which is pCKDL 23 containing spacers for arcA, fis, fnr, and hns. In this strain, the addition of fnr and fis spacers actually increased rather than decreasing the growth rate of the strain. The maximum OD fitness is dominated by hns which consistently reduces the fitness when the hns targeting spacer is present (in 14/16 cases). Targeting spacers for arcA, crp, fis, and fnr all increase the fitness (8/16, 10/16, 9/16, 3/16 respectively) when there is more than one targeting spacer present, while individually they all reduce fitness, although of the four only arcA is a significant reduction. In particular, the combination of arcA, crp and fis targeting spacers seems to have a large increase in fitness when in combination.

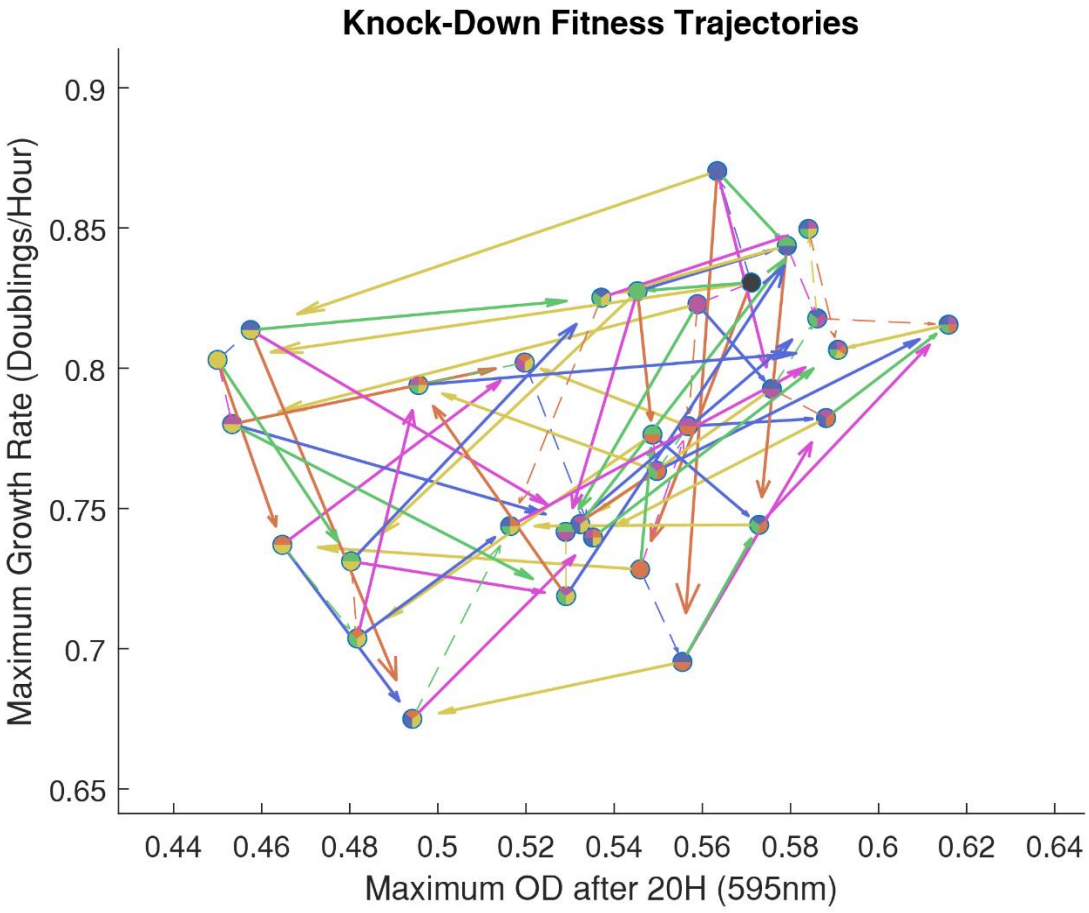



*Figure 32: The fitness landscape of pCKDL vectors grown in LB media. The non-targeting pseudo wildtype is in dark grey. Arrows represent fitness trajectories when one additional targeting spacer is added (green: arcA, blue: crp, pink: fis, orange: fnr, yellow: hns, solid: p<0.05, dashed: p>0.05). Circles are located at the median fitness measurement for each pCKDL strain, with colours representing the knockdowns present in that strain (corresponding to same colours as above).*

The fitness patterns for targeting spacers is dependent on the growth media. We performed the same growth curves with a defined media M9 which was supplemented with 0.4% Glucose as a carbon source (Figure 53). The striped pattern observed in the maximum OD reached in the LB growth curves seems to have shifted to the transition between exponential and stationary phase. We quantified the fitness measurements from these growth curves and plotted the fitness measurements and the fitness trajectories between strains which only differed by a single spacer (Figure 33). Unlike LB media, the fitness in M9 + Glucose tends to cover less of the fitness landscape, with a more linear path. It is also dominated by three extreme phenotypes, pCKDL 7 (fis, fnr, and hns spacers) and pCKDL 23 (arcA, fis, fnr, and hns spacers) have very high fitness scores while pCKDL 25 (arcA, crp, and hns spacers) has very low fitness scores. Similarly to fitness in LB media, the perturbations are non-monotonic, and many instances of cells fitness gains or losses being reversed with the addition of further targeting spacers. We decomposed the fitness for M9 media into individual figures to assess the impact of specific spacers on each fitness measurement. For growth rate fitness (Figure 54) there is a trend for fis and fnr spacers to increase the growth rate, while spacers for arcA, crp, and hns decrease the growth rate. Unlike with LB, there are no significant exceptions to these patterns. For maximum OD fitness (Figure 55), targeting spacers for fnr and fis tend to increase the maximum OD cultures reached, while targeting spacers for crp decreased the

maximum OD. Spacers for arcA and hns on the other hand where mixed, with arcA increasing maximum OD fitness once and decreasing it twice, while hns decreased maximum OD fitness once and increased it twice. We had assumed a priori that CRP would not have a strong impact in M9 with Glucose, as the presence of glucose should limit cAMP availability, a necessary co-factor for CRP activity, however we observed the opposite, with a strong negative effect when crp expression was knocked down.

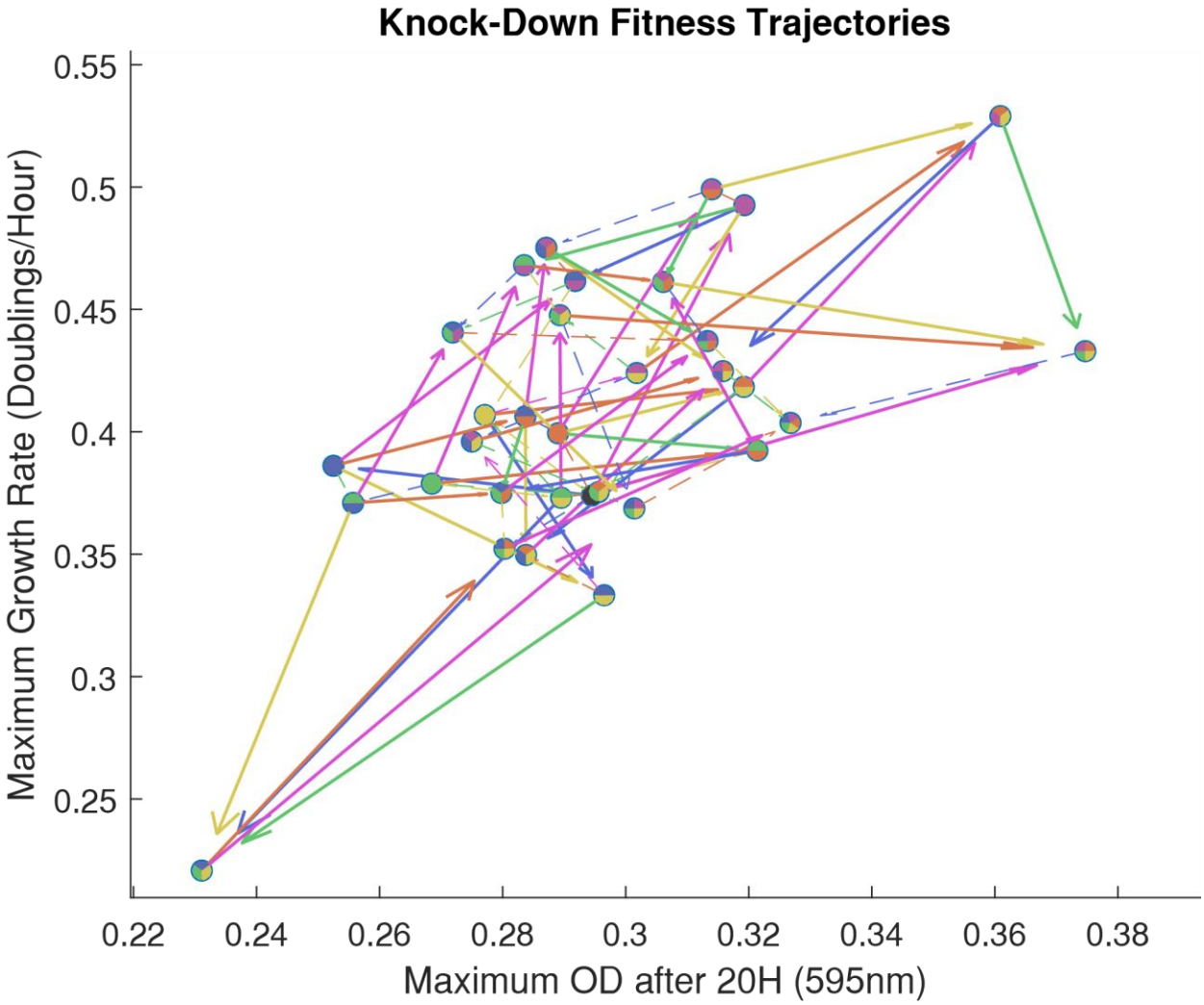



*Figure 33: The fitness landscape of pCKDL vectors grown in M9 media supplemented with 0.4% glucose. The non-targeting pseudo wildtype is in dark grey. Arrows represent fitness trajectories when one additional targeting spacer is added (green: arcA, blue: crp, pink: fis, orange: fnr, yellow: hns, solid: p<0.05, dashed: p>0.05). Circles are located at the median fitness measurement for each pCKDL strain, with colours representing the knockdowns present in that strain (corresponding to same colours as above).*

Finally, we measured growth curves in M9 media supplemented with 0.4% Lactose as a less preferred carbon source (Figure 56). Unlike LB or M9 with Glucose, there are no obvious patterns from the comparisons of growth curves directly, although combinations of fis and fnr (pCKDL 6, 7, 14, 15, 22, 23, 30, 31) and combinations of fis and arcA (pCKDL 20,21,22,23,28,29,30,31) all tend to grow faster and to a higher maximum OD. We quantified the fitness metrics and plotted them along with the trajectories between strains (Figure 34) as before with LB and M9 + Glucose. Unlike LB or M9 + Glucose, when grown with Lactose as a carbon source strains with a high fitness for one measure tended to have a lower fitness

for the other (such as pCKDL 6 and pCKDL 23). Examining the individual fitness scores, for exponential growth rate (Figure 57) only the fis spacer increased growth rate. While arcA, crp and hns spacers decreased growth rate, this was only significant in twice, once and once respectively. Additionally fnr had one significant increase in growth rate and one significant decrease in growth rate. The maximum OD fitness (Figure 58) showed significant increases with the addition of fnr or hns targeting spacers, and while crp also increased the maximum OD fitness, it did so only once significantly. The arcA targeting spacer was the only spacer to consistently decrease maximum OD fitness in lactose media, although it only did so in 2 cases. Finally, the fis targeting spacer increased fitness increased fitness in 3 cases and decreased fitness in 2 cases. Again, we were surprised that in lactose media crp didn't have a stronger effect, as canonically crp is required for expression of the lac operon and therefore import and metabolism of lactose. This may be a consequence of our perturbation system, in which sufficient levels of crp or lacY are present from pre-induction to maintain lactose import and metabolism.

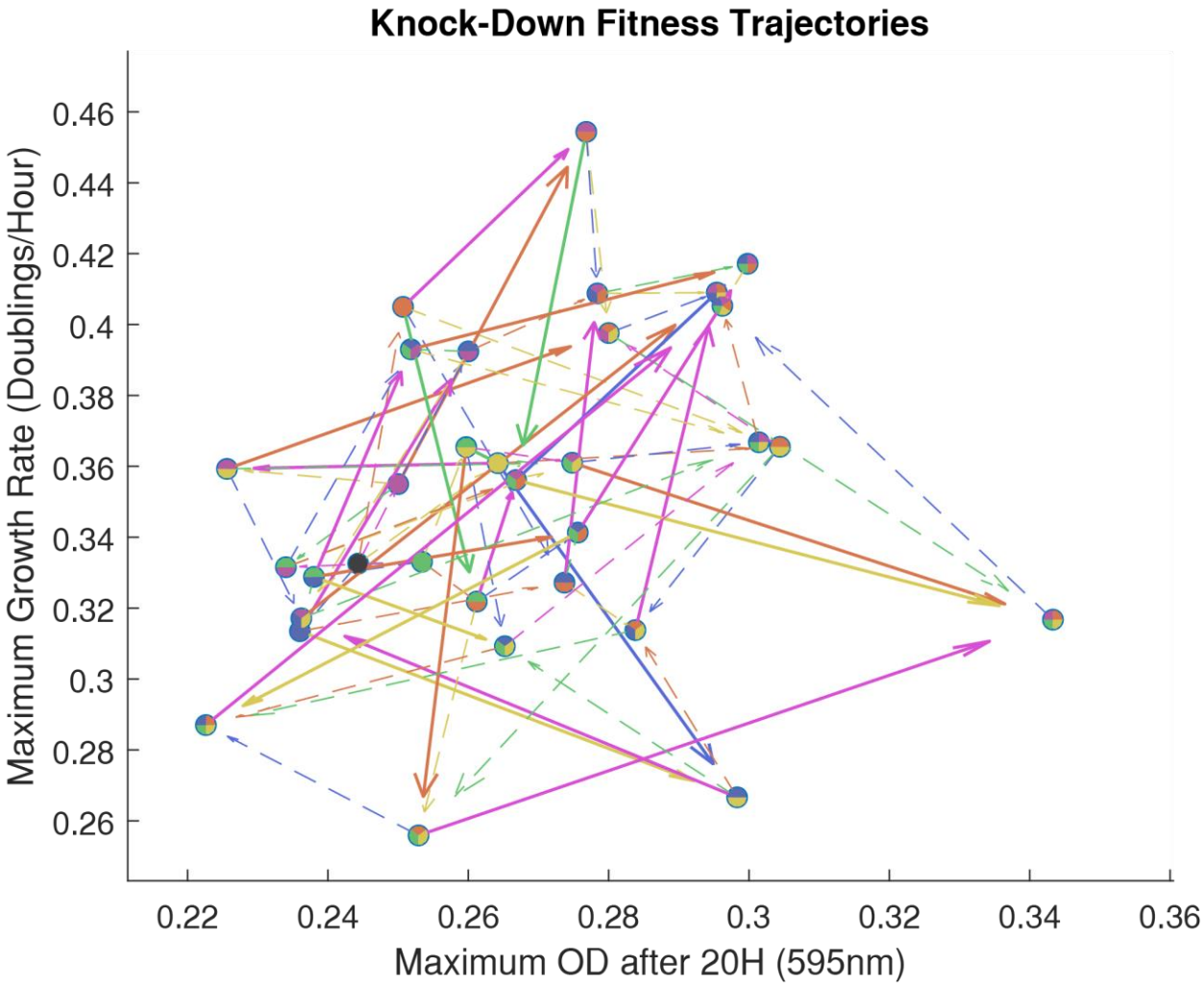



*Figure 34: The fitness landscape of pCKDL vectors grown in M9 media supplemented with 0.4% glucose. The non-targeting pseudo wildtype is in dark grey. Arrows represent fitness trajectories when one additional targeting spacer is added (green: arcA, blue: crp, pink: fis, orange: fnr, yellow: hns, solid: $p<0.05$, dashed: $p>0.05$). Circles are located at the median fitness measurement for each pCKDL strain, with colours representing the knockdowns present in that strain (corresponding to same colours as above).*

***High order epistasis is present in the fitness landscapes***

Given that the fitness trajectories were non-monotonic and implied epistasis we compared the fitness measurements of the double perturbations to the single perturbations to see if this was the case. Due to the combinatorial nature of our perturbation library, there are 8 different backgrounds for every pair of perturbations. For each pair of perturbations in each background, we normalized the fitness such that the fitness of the reference strain (the 8 strains not containing either of the pair of targeting spacers) had a fitness of 1. We then calculated the predicted fitness of the double perturbation assuming no epistasis by multiplying the fitness of each of the single perturbations. We then calculated the difference between the expected (no epistasis) and the observed fitness for each pair of perturbations. We found that this was not consistent between all 8 backgrounds. We tested if the epistasis we observed was influenced by the presence or absence of another targeting spacer in the reference strain, or if this was just a result of experimental noise. For each of our 10 pairs of genes, we split the references strains depending on either they had a targeting or non-targeting spacer for each of the 3 remaining genes, for all 6 fitness measurements. This resulted in 180 samples, in which 42 showed significantly different epistasis between reference strains with or without an additional targeting spacer (Figure 59). While some of the p values are only slightly below the threshold for significance and maybe the result of false positives, others show very strong effects. Strikingly there are cases when the presence of a third perturbation inverts the sign of the epistasis interaction between two perturbations. For example, the maximum OD fitness in LB media for hns and fnr has positive epistasis in the absence of an arcA spacer, but negative epistasis in its presence ($p = 6.7*10^{-11}$). Likewise the growth rate fitness in M9 + Glucose fitness for fis and crp shows negative epistasis in the absence of a hns spacer, but positive epistasis in its presence ($p = 0.0019$). These interactions were graphed with nodes representing the targeting spacers and edges representing the epistasis (in absence of a third targeting spacer). Arrows from the nodes to the edges indicate how the presence of that targeting spacer changes the observed epistasis (Figure 60). In many cases, when a third spacer influenced the epistasis of a pair of perturbations, the inverse would also be true, in that each spacer in that pair would influence the epistasis between the remaining two spacers. In cases when 3 genes all influenced the interaction between each other in all combinations, those three genes are grouped in a dashed line. We found 8 of these groups in total, indicating that these influences are internally consistent.

To explore the possibility of higher order interactions more thoroughly, we applied linear regression to our fitness measurements to calculate the epistatic coefficients, up to and including for all 5 perturbations combined (n = 5) [55]. We then sought to systematically eliminate higher order coefficients, and recalculated the epistatic terms (r = 1 to 5). This was still done with all the fitness measurements for all 32 strains, but calculated a reduced number of coefficients. For example, when r = 1 only coefficients for the reference strain and the single perturbations are calculated, and when r = 2 the coefficients for all pairwise perturbations are calculated in addition to the coefficients for r = 1. The coefficients for r = 1 to 5 are then used to calculate an expected fitness for each strain and this fitness was normalized by the observed fitness for each strain (Figure 35). Regression coefficients up to the $4^{th}$ order (r = 4) were needed to accurately predict the fitness for growth rate for all conditions, and coefficients up to the $5^{th}$ order (r = 5) were need for maximum OD fitness, indicating that there are higher order epistasis interactions between the master transcriptional regulators. We plotted the regression coefficients (r = 1 to 4) against the regression coefficients (r = 5) for both growth rate (Figure 61) and maximum OD fitness (Figure 62). As expected, when r = 4 the regression coefficients for growth rate fitness collapse into a line. This

however does not occur with maximum OD fitness, and the correlation between regression coefficients actually decreases when r is increased from 3 to 4. We can also see that in every case except r = 4 for growth rate, there are instances when coefficients are found within the blue squares. This is important as it changes the interpretation of the epistasis between two genes.

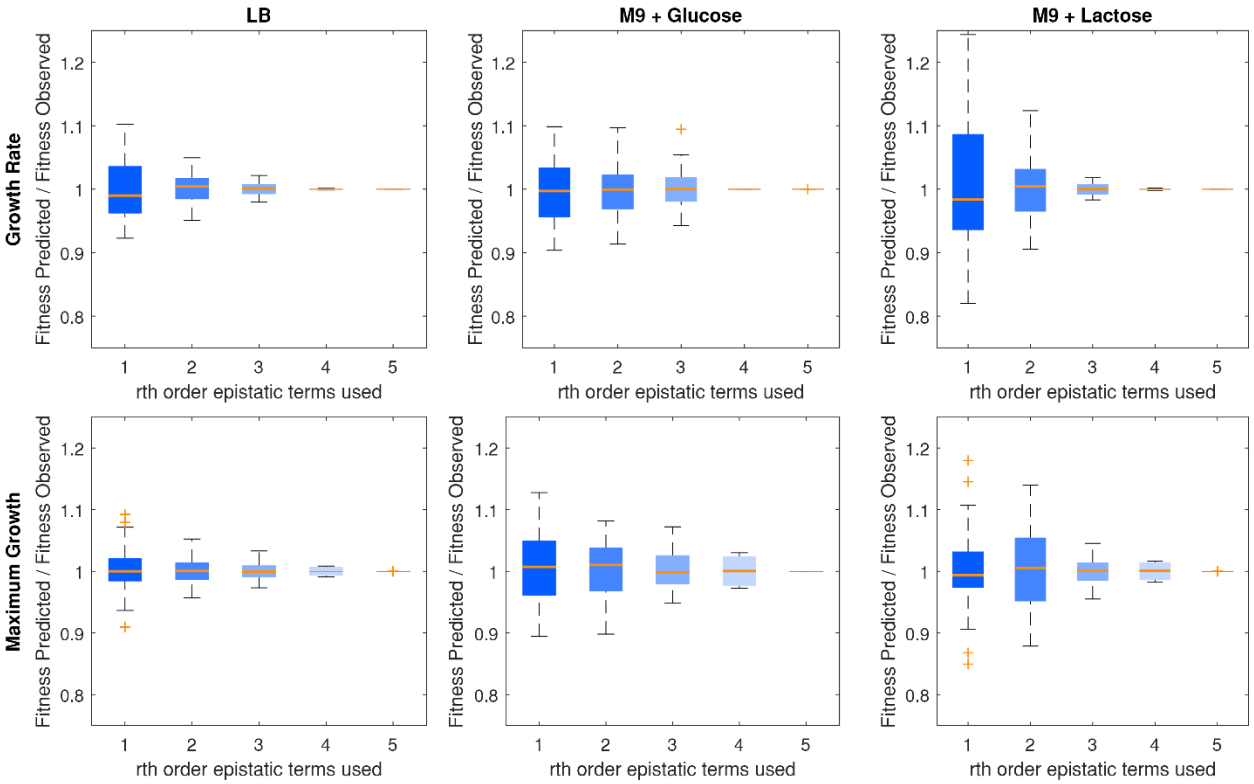



*Figure 35: Predictability of fitness from regression coefficients corresponding to $r^{th}$ order epistasis. Regression coefficients are computed using observed fitness measurements for all 32 CKDL strains. Coefficients computed are limited to the $r^{th}$ order, such that $1^{st}$ order represents only coefficients for unperturbed and single gene perturbations, $2^{nd}$ order additionally includes coefficients for pairwise perturbations, $3^{rd}$ order additionally includes triple perturbation coefficients, etc. These regression coefficients are then used to predict the fitness of all 32 CKDL strains. Box plots represent the predicted fitness divided by the observed fitness for all strains. Boxes represent the $25^{th}$ and $75^{th}$ percentile, the orange line representing the median of the data, and whiskers represent the limits of the data not considered outliers. Outliers are plotted individually as orange '+'.*

In this case, we have a clear reference strain in the non-targeting pCKDL 0 vector. However there is also the opportunity to compute the epistasis in terms of [-1 1] instead of [0 1]. This could allow us to calculate a background-averaged epistasis and compare if it is able to fit the data with fewer higher order terms [55]. Additionally, while we have compared the predicted fitness to the observed fitness with different orders of epistasis removed, we could also measure the prediction error in terms of the total number of epistatic coefficiences required [137]. This involves removing coefficients that weakly contribute to predictive power one by one regardless of their order. Epistatic coefficients can systematically be set to zero with each step reducing the smallest non-zero coefficient. This analysis is still on-going while we seek to find the most appropriate method for calculating epistasis.

## 3.4 Transcriptional Profiles of pCKDL Perturbations

***RNAtag-Seq allows multiplexed RNA profiling comparable to traditional RNA-Seq***

To investigate the effect of perturbing global transcriptional regulators has on gene expression, we performed RNAseq on our pCKDL vectors. Our target genes are responsible for directly regulating a large number of genes. We use a technique called RNAtag-Seq to multiplex our samples. This involves ligating an RNA tag with a specific barcode to RNA for each sample. As a result, all the RNA from a single sample will carry a unique barcode. The samples can then be pooled for ribosomal RNA depletion, reverse transcription, and amplification. To ensure that RNAtag-Seq provided similar results to previously published experiments, we performed the RNAtag-Seq protocol on RNA from MG1655 grown in log phase (to an OD of 0.15) in M9 + Glucose media. After sequencing the cDNA, we aligned reads from our RNAtag-Seq experiment as well as GEO database series GSE61327, GSE65711, GSE66481, GSE48324, and GSE48829 to the MG1655 genome using RockHopper. We then plotted the gene expression for each GSE experiment against our RNAtag-Seq data and calculated the correlation between experiments (Figure 36). All the GSE experiments we used were exponential phase MG1655 grown in M9 + Glucose. The RNAtag-Seq was highly correlated with all of the GSE experiments except GSE66481, which had a similar correlation to our data as to the other GSE experiments. A key difference with GSE66481 is that RNA was extracted in acid conditions with a pH of 5. The RNAtag-Seq protocol gave comparable results to previously published bulk RNAseq results.

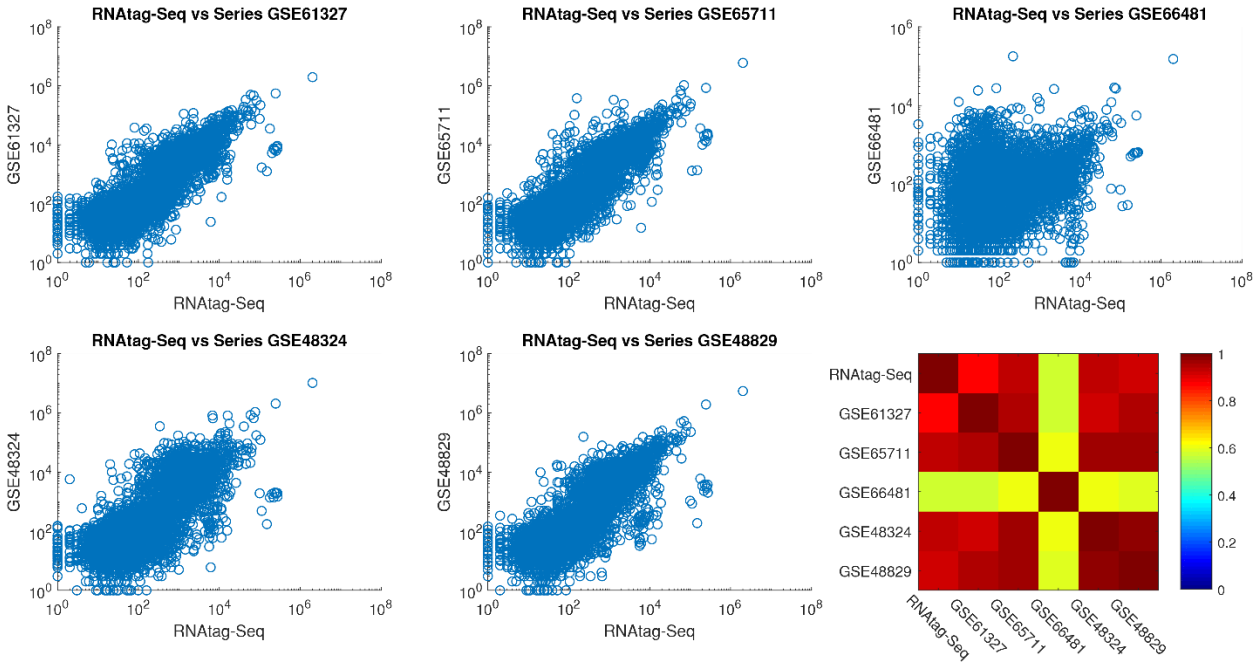



*Figure 36: Comparison of RNAtag-Seq results against bulk RNASeq experiments from GEO database. RNAtag-Seq was performed on RNA extracted from MG1655 grown in M9 medium + 0.4% Glucose until an OD of 0.15. Fastq files from the RNAtag-Seq and series GSE61327, GSE65711, GSE66481, GSE48324, and GSE48829 were aligned to E. coli MG1655 genome. The correlation between experiments is shown.*

***RNA sequencing shows perturbations correspond to distinct principle components***

We began by preparing RNA for all 32 pCKDL grown in LB media for 1.5 hours or 3.5 hours, these correspond to mid exponential phase (OD ~0.3) and early stationary phase (OD ~1.3) respectively. We extracted RNA using Qiagen RNAProtect and RNeasy Kits and performed qRT-PCR on the samples to ensure that the pCKDL vectors were functioning (Figure 63). We saw a noticeable shift in the Cq for each gene in samples which contained a targeting spacer for that gene, despite all RNA samples having a similar Cq for a Spiked RNA of a known concentration. Before proceeding with RNAtag-Seq, we quantified the quality of the RNA sample extracted (Figure 64) using an Agilent TapeStation. The majority of samples had an RNA integrity number (RIN) above 8, although any samples with an RIN below 7 were repeated. We also recorded the OD of each sample before RNA extraction using 200 µL of culture in a microtitre plate (Figure 65). While the OD of samples from 1.5 hours have a fairly tight distribution, this begins to spread by 3.5 hours, as expected by the different growth rates observed for the various pCKDL strains. RNAtag-Seq was done on all 64 RNA samples and sequenced using an Illumina NextSeq. Sequencing Reads were aligned to the MG1655 genome using RockHopper which converts them into reads per kilobase per million mapped reads, although it is normalized by the upper quartile of gene expression rather than by total mapped reads to improve robustness. We compared the number of detected reads for strains with the targeting spacer or the non-targeting spacer for each of our targeted genes (Figure 66). We found that gene expression of our targeted genes was nearly abolished by the CRISPR-dCas9. Additionally, we did not detect any strains which managed to successfully escape repression from the CRISPR-dCas9.

We performed Principle Component Analysis (PCA) on our gene expression data and found that the first 10 Principle Components (PC) account for only 58.94% of the variance in the data. The first 22 Principle Components explain 75.54% of the Data and the first 48 Principle Components are needed to explain 95.33% of the variance in the Data (Figure 67). We then plotted the principle component scores for the first 8 principle components and colored each point by various qualities of the sample to determine if the certain principle components were associated with specific features. Since the first principle component is often reported to be associated with growth phase, we began with OD (Figure 68). We clearly have a separation in the first principle component between samples taken in exponential phase and those taken in early stationary phase. We also see a clustering of the exponential phase strains in the center of the 3$^{rd}$ and 4$^{th}$ principle components, with the stationary phase strains comprising 4 protrusions in opposite directions. We then checked the principle component scores against each of our 5 perturbations: arcA (Figure 69), crp (Figure 70), fis (Figure 71), fnr (Figure 72), and hns (Figure 73). ArcA spacers strongly separate out on the 5$^{th}$ principle component, but they also form clusters of 16 in the 1$^{st}$ and 2$^{nd}$ principle components, and clusters of 4 in the 3$^{rd}$ and 4$^{th}$ PCs. Strains containing a crp targeting spacer cleanly separate on the 3$^{rd}$ principle component, though they also tend to be separated on the positive side of the 1$^{st}$ principle component for each separate growth phase group. Fis spacers cleanly separate on the 6$^{th}$ PC and weakly on the 7$^{th}$ PC while FNR is the opposite in strongly separating on the 7$^{th}$ PC and weakly on the 6$^{th}$. Finally hns spacers separate distinctively in the 4$^{th}$ principle component. We also check against batch effects (Figure 74) as they are known to have a strong impact on expression data. We clearly see individual batches clustering in the first two principle components, which also helps to explain the clustering of arcA strains in the first two principle components as well, as all 16 strains in each batch contain the same spacer for the arcA position as a consequence of how strains are ordered. With the identification of batches clustering in the 1$^{st}$ and 2$^{nd}$ PCs, we decided to see if the second principle

component could be related to biases from amplification. As all pooled batches began with approximately the same quantity of RNA, differences in RNA cycles required to reach a concentration sufficient for sequencing would indicate biases in the expression data introduced from either PCR amplification, RT-efficiency, or Ligation Efficiency. We are unable to disentangle those three sources of error, but we clearly see that the 2nd principle component does correspond to the number of PCR cycles used (Figure 75), indicating that these likely have a strong effect on the data. We summarized the principle components dominated by a particular perturbation (Figure 37). PC3 is associated with crp, PC4 is associated with hns, PC5 with arcA, PC6 with fis and PC7 with fnr. This ranking does not seem to correspond with either their hierarchy in the regulatory network, nor their total number of regulatees either direct or indirect.

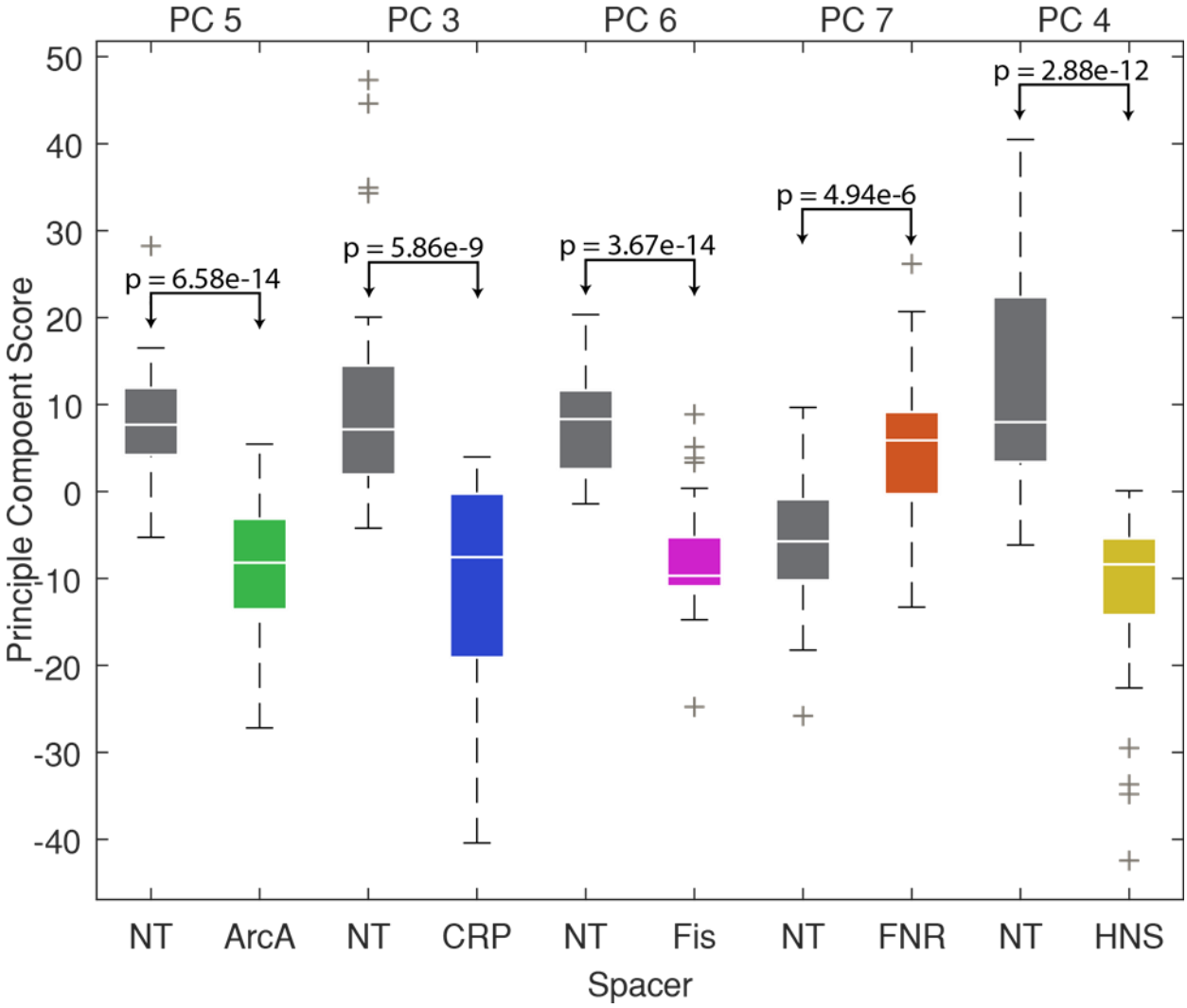



*Figure 37: Association of targeting spacers to specific principle components. The principle component scores for pCKDL strains that contain either a targeting spacer (coloured boxes; green: arcA, blue: crp, pink: fis, orange: fnr, yellow: hns) or a non-targeting (NT) spacer (grey boxes). The principle component (PC) associated with each perturbation is indicated above the box plot. Boxes represent the 25th and 75th percentile, the white line representing the median of the data, and whiskers represent the limits of the data not considered outliers. Outliers are plotted individually as grey '+'.*

***Principle components are consistent with regulatees for each perturbation***

We mapped genes which are directly or indirectly regulated by our global transcription factors to the gene's contribution to the associated principle component. We observed a strong enrichment of the location of genes directly or indirectly regulated by ArcA (Figure 38) to genes strongly contributing to the 5$^{th}$ principle component, CRP (Figure 76) to genes strongly contributing to the 3$^{rd}$ principle component, Fis (Figure 77) to genes strongly contributing to the 6$^{th}$ principle component, and FNR (Figure 78) to genes strongly contributing to the 7$^{th}$ principle component. There also seemed to be enrichment of genes directly regulated by HNS (Figure 79) to genes contribution to the 4$^{th}$ principle component, but it was not as distinct as the other transcription factors, nor did it seem to extent to indirectly regulated genes.

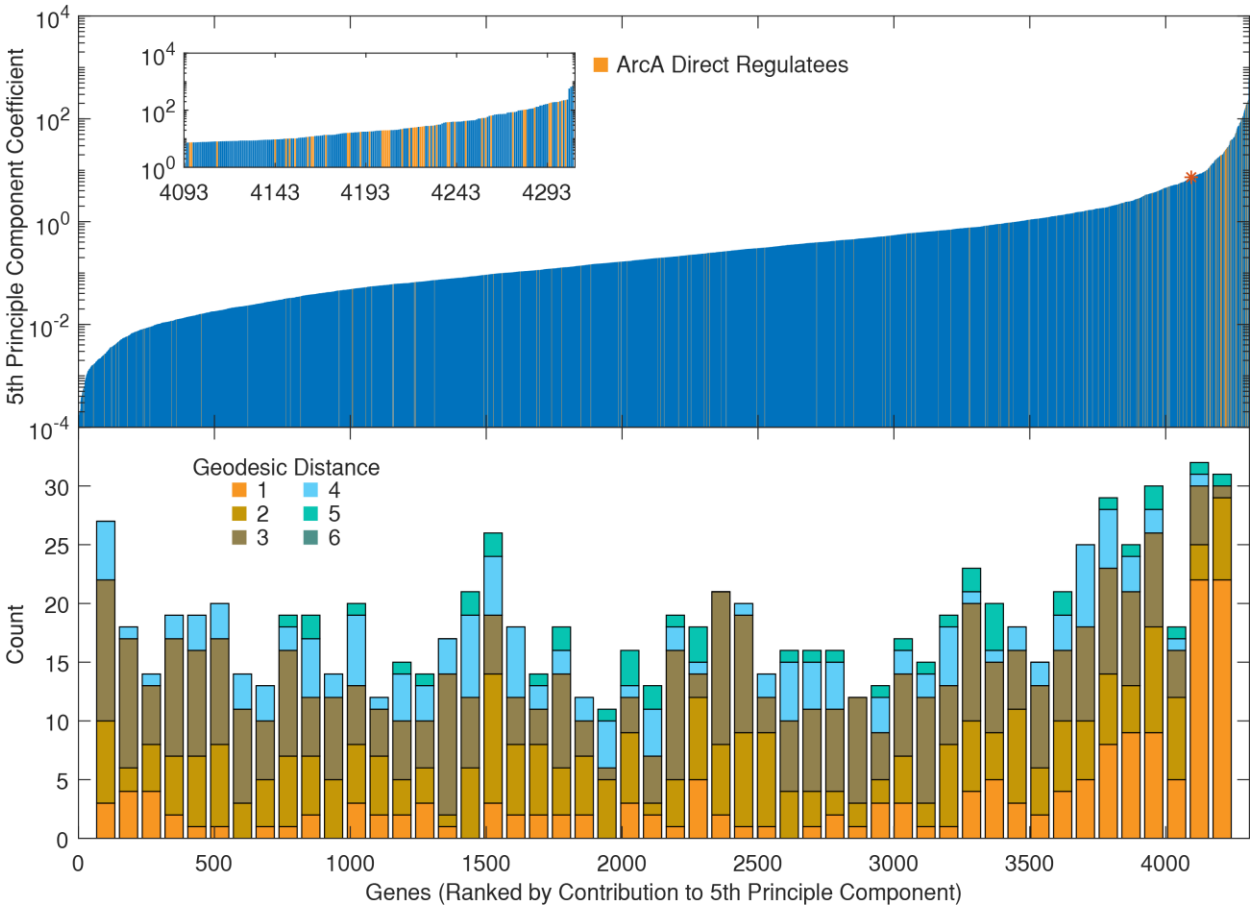



*Figure 38: Mapping of direct and indirect regulatees of ArcA onto genes sorted by their contribution to the 5$^{th}$ Principle Component. Top: Genes' contribution to the 5$^{th}$ principle component. Genes highlighted in yellow are directly regulated by ArcA. The star on the curve represents contributions above 1% of the maximum. Insert is an expanded view of all the genes above this threshold. Bottom: Genes are separated into 50 equal sized bins. The number of genes in each bin which are directly regulated by ArcA (Geodesic Distance of 1) are shown in yellow. The number of genes in each bin which are indirectly regulated by ArcA (Geodesic Distance 2 to 6) are also shown.*

***The first principle component corresponds to sigma factors and growth state***

To determine if the proportion of all sigma factor reads corresponded to a principle component, we tested if the sigma factor proportion was correlated to principle components (Figure 80). We saw a significant correlation between RpoD, FecI, RpoE, RpoH, and RpoN, and a significant anti-correlation of

RpoS, with the first principle component. We also saw a weak anti-correlation of RpoD and RpoN with the 5th principle component. This leads us to believe that changes in sigma factor ratios are largely driving the first principle component and therefore growth phase, while the activity of the global regulators seems to be largely independent from sigma factors.

As our sequencing data indicated that the first principle component corresponded to growth state (by the OD of the sample when RNA was extracted), we tested to see if our genes would form two distinct anti-correlated clusters when sorted by the first principle component, similar to the many microbe microarray dataset [42] [132]. We found the same pattern occurred in our RNA-sequencing data, however it was mirrored compared to the micro array data (Figure 39). We then checked the correlation between the first principle component scores and the optical density of the samples, and found that in our principle component analysis, they were anti-correlated while in microarray data they are correlated. This results in the mirrored effect that we observe. We checked to see if the same genes were found in each cluster, and had a fairly good overlap between data sets (Figure 81). Given that microarray and RNA-seq do not strongly correlate [138] [139] [140] [141] we were satisfied with this consistency.

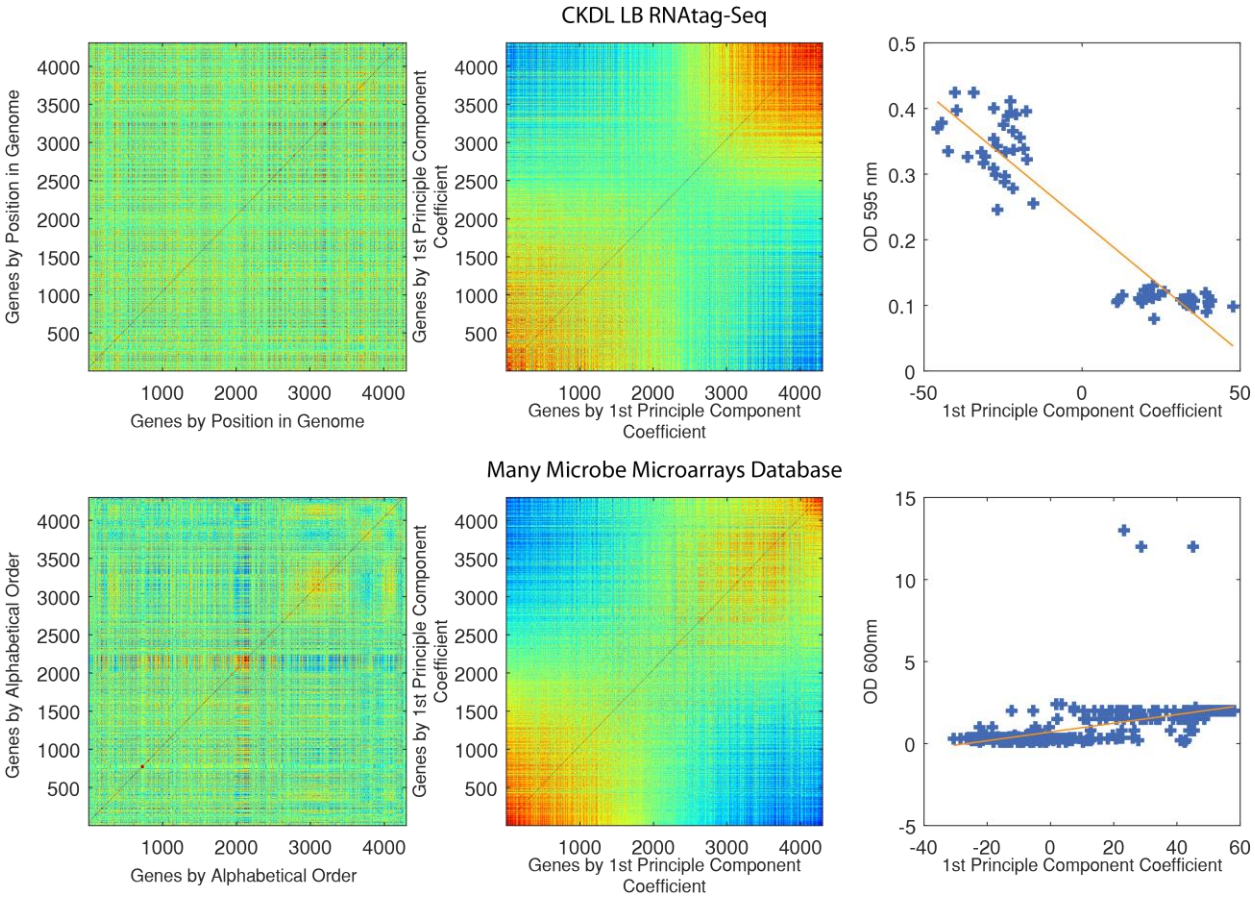



*Figure 39: Comparison of pCKDL data with Many Microbe Microarrays Database. Correlation of gene expression is shown in pCKDL and Many Micro Microarray Database (M3D) with genes in their default order, either by genomic position or alphabetically respectively. Genes are then sorted by their contribution to the first principle component resulting in two anti-correlated clusters of highly correlated genes. The correlation between each datasets first principle component and the optical density of the samples is also plotted.*

***Maximum growth but not growth rate maps to principle components***

Finally, we checked to see if any of our principle components corresponded with our fitness measurements (Figure 82). We found that the maximum OD strongly correlated with the 4th principle component ($r = 0.7086$ $p = 5.6 \times 10^{-6}$), the same component that was associated with hns. This is consistent with the striped pattern in the LB growth curves (Figure 50). As such it is likely that this association will not hold true with the sequencing results from M9 media, as this pattern is not observed in their growth curves. The maximum OD fitness was only weakly correlated with the 6th and 8th components after the 4th component. We didn't find the maximum growth rate fitness to significantly correlate with any of our first 7 principle components, and only found weak correlations with the 18th and 19th principle components. This likely indicates that the growth rate fitness is not captured by our transcriptome analysis which may be more closely tied to ribosome synthesis and proteome demands [142] [143].

***High order epistasis is consistent across clusters of gene expression***

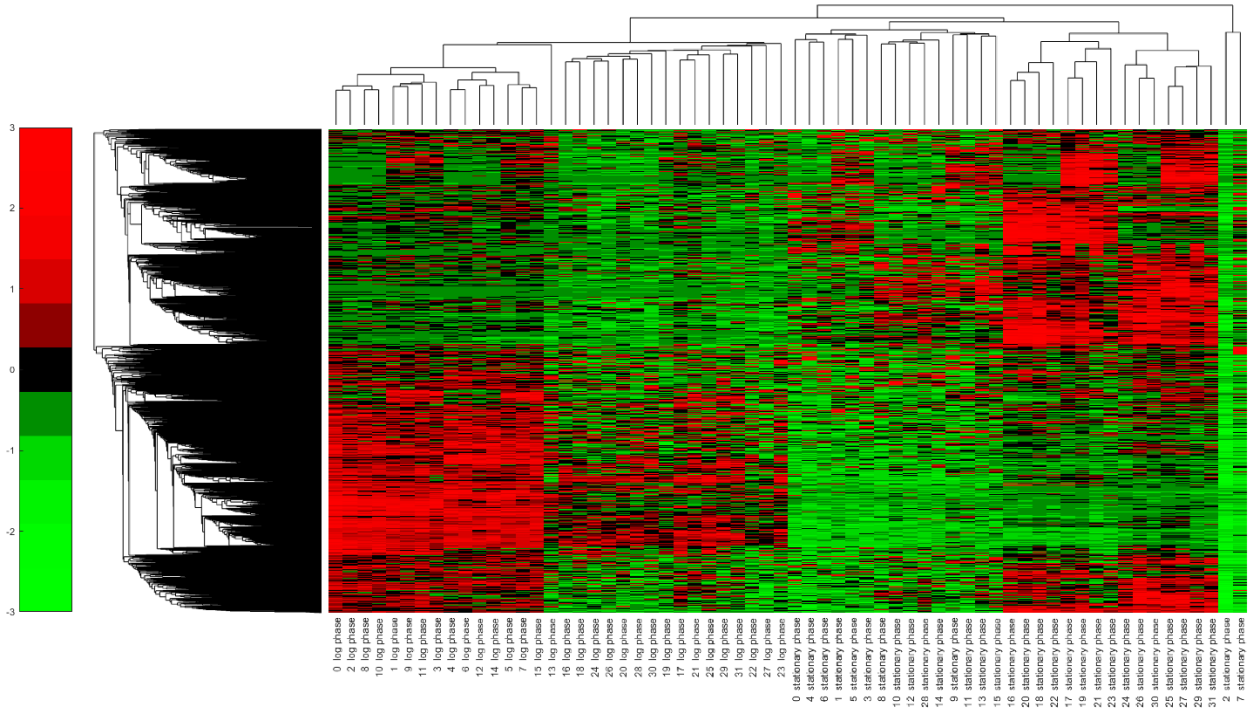


*Figure 40: Clustergram of pCKDL RNA-seq data from LB culures. The Z-score of the transcriptional data was clustered using a clustergram algorithm in Matlab. This clusters both the experiments (pCKDL samples) and the measurements (Gene Z-scores). The perturbations in each pCKDL corresponds to the 5-bit binary vector of the sample number. The data is clustered into log phase on the left, and exponential phase on the right. Within each of these, wild-type (WT) arcA (0-15) is on the left and knock-down (KD) arcA (16-31) is on the right. These can further be decomposed, for example within the left most cluster of 16, the first 8 correspond to WT fis (0,1,2,3,8,9,10,11), while the second 8 correspond to KD fis (4,5,6,7,12,13,14,15). Within these, the first four correspond to WT hns (0,2,8,10) and the second four to KD hns (1,3,9,11). Within those are WT crp (0,2) and KD crp (8,10), and finally within those is WT fnr (0) and KD fnr (2). It is important to note that this order is not conserved for all branches.*

Our principle component analysis indicates that the primary effects of each of the global transcriptional regulators on gene expression are independent. However this only explains less than 60% of the variance in our data, and is contrary to the higher order epistasis we observed in our fitness measurements. We therefore attempted to examine the RNA-sequencing data with cluster analysis. We

first filtered all genes with low variance, low absolute expression levels, and low entropy [144], however this only eliminated approximately ¼ of the genes. We used the GAP method [145] to determine the optimal number of clusters with both Linkage and Kmeans methods. We found that 5 clusters would have the optimal trade-off between having a few number of clusters for both hierarchical and K-means methods of clustering (Figure 83). We then clustered the genes using both a Hierarchical algorithm and K-means [146], as well as producing a clustergram which performs hierarchical clustering on our expression data and displays a dendrogram and heatmap (Figure 40). The hierarchical tree is generated using the Euclidean distance metric and average linkage. We found (4/25/15/1271/1979) genes in each hierarchical cluster and (1118/403/670/520/583) genes in each K-means cluster.

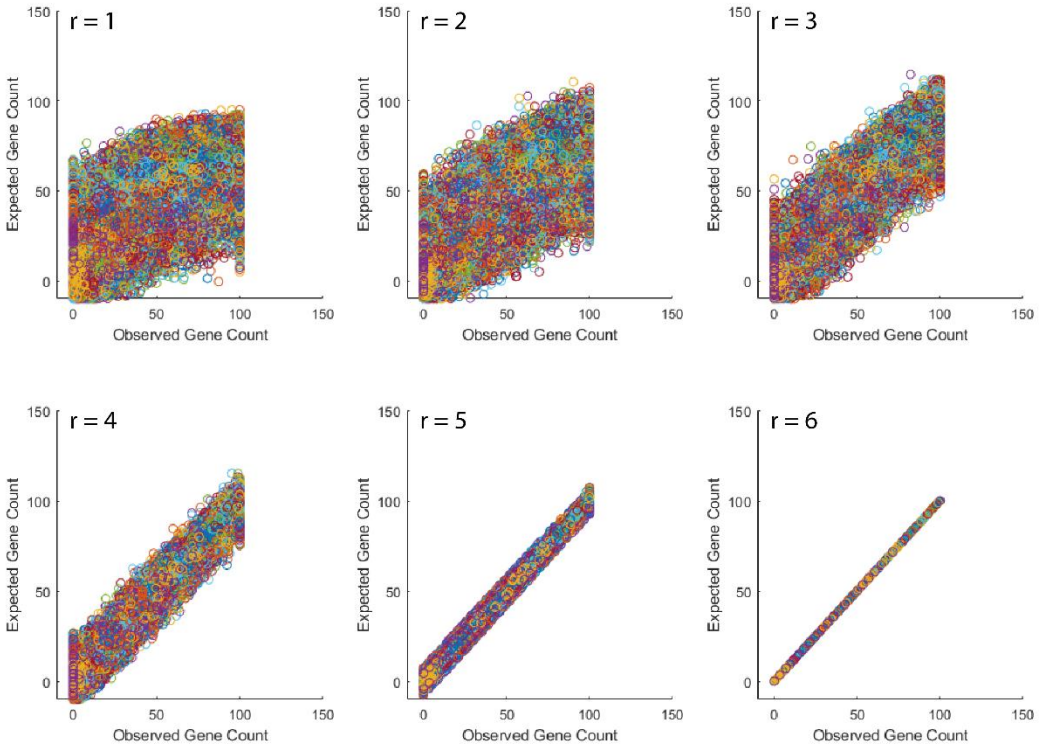



*Figure 41: Prediction of gene expression using r-th order epistasis coefficients. Epistasis coefficients are determined by linear regression, where r = 6 calculates all possible epistasis coefficients and therefore perfectly matches the observed data. As r decreases, fewer coefficients are calculated. All gene counts are scaled to a range between 0-100 for comparison.*

We found clear patterns in the clustergram which corresponded to our growth state and the genetic perturbations. The growth state was clearly separated between exponential phase samples and stationary phase samples. This also corresponded to our first principle component in the PCA analysis. The second layer for each of these separated by arcA perturbations. This separation is harder to attribute completely to arcA perturbations, as our PCA revealed that the second component was associated with each RNA-seq batch, and specifically PCR amplification. These batches also happen to split along arcA perturbations. Within these layers, the next separation observed was to be fis within the non-perturbed (0) arcA and hns within the perturbed (1) arcA during exponential phase but crp during both stationary phase groups. We calculated the epistatic coefficients of each gene as we had done previously for our

fitness measurements [55]. Again, we attempted to calculate only the coefficients for r order terms (such that r = 1 is no epistasis, r = 2 is pair-wise epistasis and so forth) (Figure 41). Here, we included growth state as another term, with exponential phase as 0, and stationary phase as 1. We can clearly see the predicted gene count spreads as we reduce the number of epistatic coefficients. Additionally, when r is < 4, the trend begins to curve downward, underestimating the gene count. This is likely due to nonlinearities in our data that are not accounted for in our linear regression model. As the nonlinearities increase, the data will progressively bend [57]. However the increasing spread in the expected gene count is the result of increasing epistasis. We separated the genes by their cluster and repeated the linear regression, but we did not find any significant change in this pattern, with all clusters converging at approximately the same rate.

***Gene clusters show patterns of multi-level complex logic***

Plotting the K-mean centroids showed some regular patterns (Figure 42). For example in the first cluster, the left-side have higher centroids than the right half. These genes are expressed in exponential phase but not stationary phase. The inverse is true for the fourth cluster. Perturbations also follow a distinct pattern. HNS perturbations are every other strain and result in a very jagged pattern (such as cluster 2). FNR perturbations are every two strains. Fis perturbations are every 4 strains. CRP perturbations are every 8 strains (such as cluster 5). ArcA perturbations are every 16 strains. We can see both patterns involving the perturbations in the clusters, but also that the amplitude of the peaks varies in different conditions.

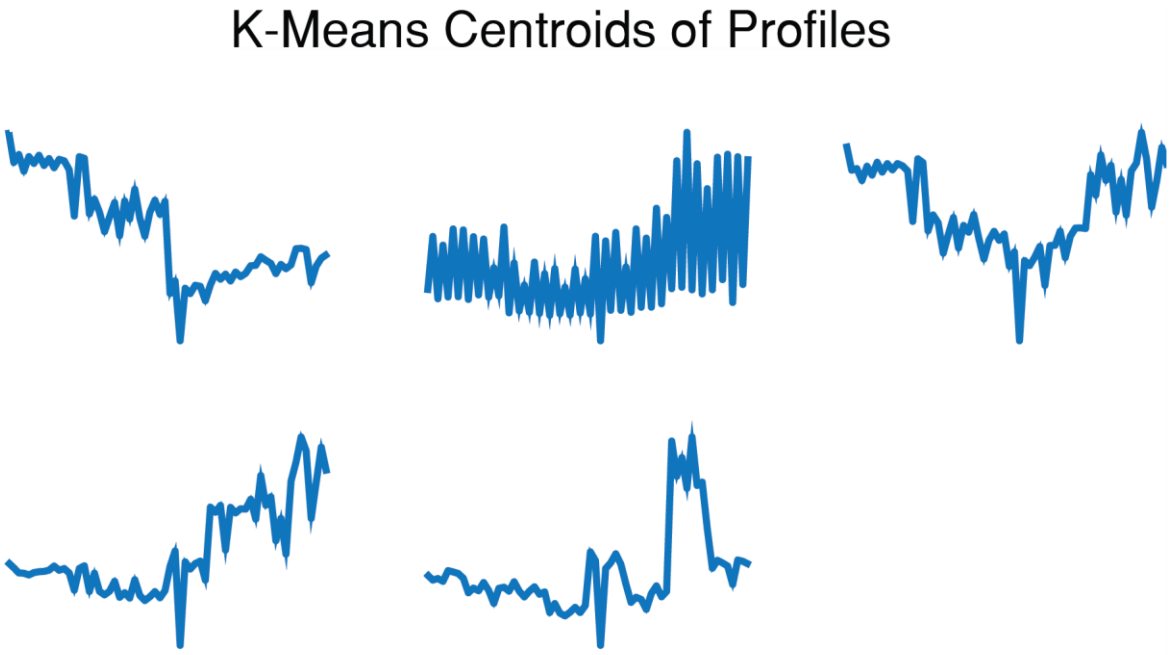



*Figure 42: K-mean centroids of pCKDL strains. Strains are ordered pCKDL 0 to 31 in exponential phase and then 0 to 31 in stationary phase. The pattern represents the mean expression for all genes within the cluster, with peaks indicating a higher expression and valleys representing lower expression.*

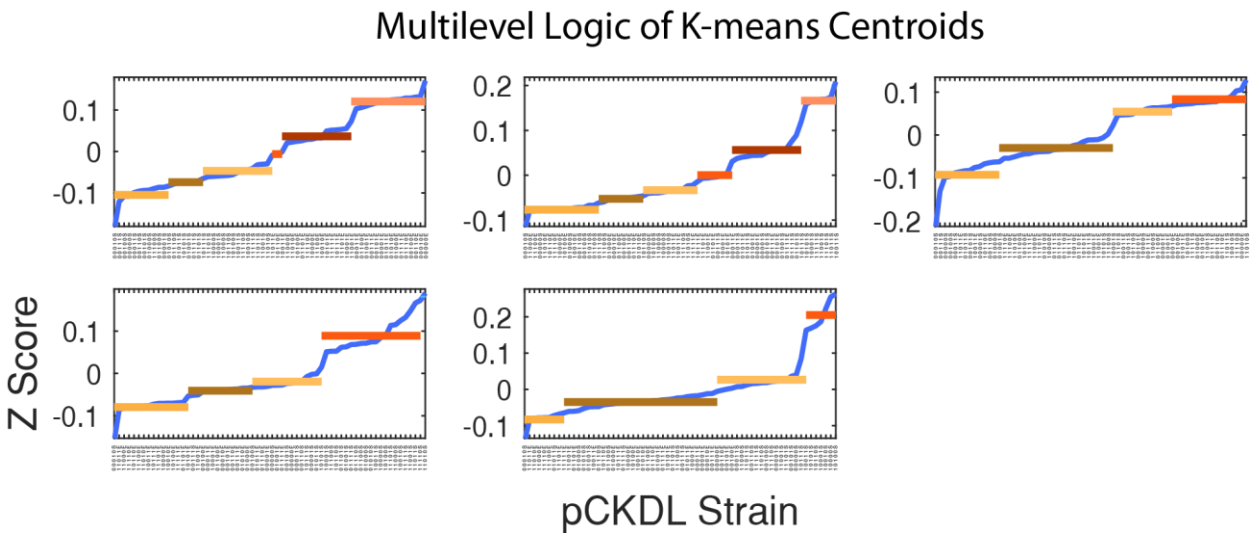



*Figure 43: Multi level logic of K-mean centroids. Strains are sorted by their centroid. Plateaus are fitted by finding local maxima of the derivative of the curve, after smoothing. These local maxima are the locations of the steps. The perturbation code (x-axis) gives the logic which corresponds to each level of gene expression in each K-means cluster.*

We turned to a logical model to attempt to explain the higher order epistasis. We therefore sorted the K-mean centroids to see if they appeared to have Boolean step functions (with genes either on or off). Instead, it appeared to have multiple steps, with genes capable of being expressed at a variety of levels. To determine these levels, we smoothed the sorted K-mean Centroids and took the derivate. We again smoothed this line and identified the local maxima. The smoothing was done to minimize the number of steps. A single step is between 2 local maxima in this smoothed derivate. We then looked at the perturbation sequence for the sorted K-Mean plots. This gives us a multi-level truth table to decipher the regulatory logic within the clusters. While some logic functions can be deduced directly from the graphs, we are still working the best way to find the minimal logic expression to describe the clusters.

***Removing principle components for growth and batch effects affects clustering but not epistasis***

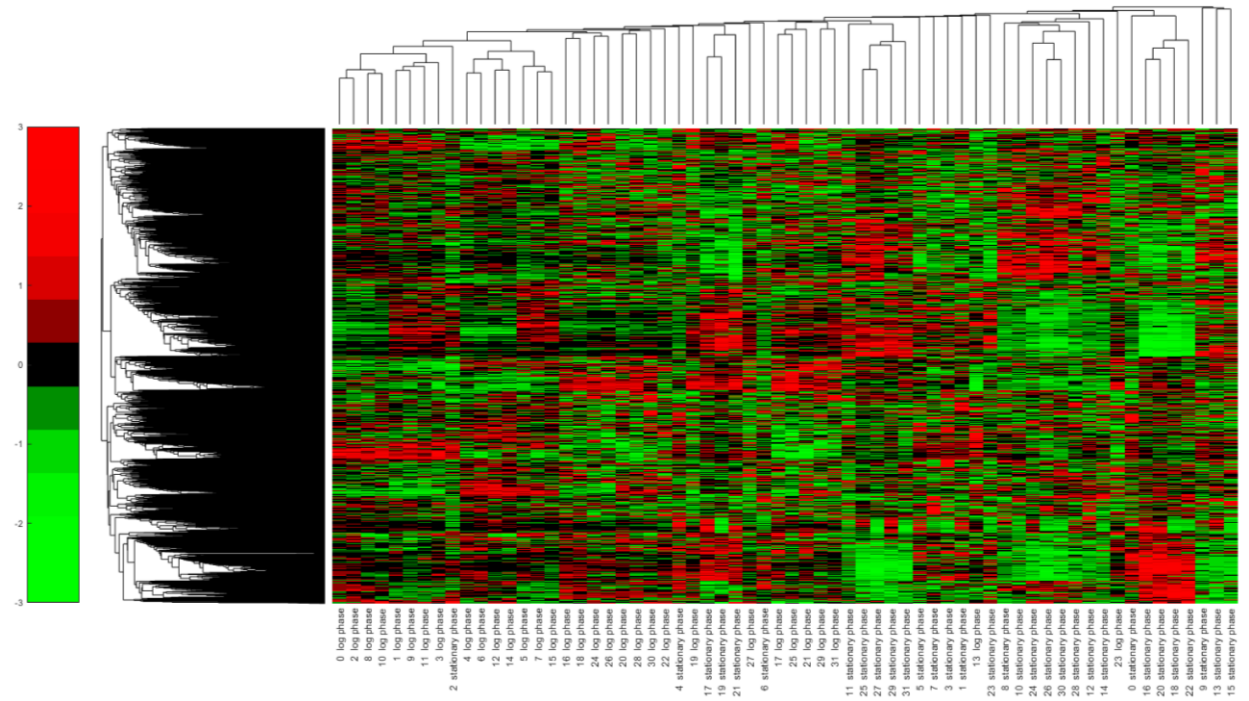


*Figure 44: Clustergram of pCKDL RNA-seq data from LB culures after the first 2 principle components from principle component analysis are removed from the data. The Z-score of the transcriptional data was clustered using a clustergram algorithm in Matlab. This clusters both the experiments (pCKDL samples) and the measurements (Gene Z-scores). The perturbations in each pCKDL corresponds to the 5-bit binary vector of the sample number. The first two principle components were found to be associated with growth state ($OD_{595nm}$) and batch effects respectively. Their removal leads to a much more nuanced clustering of genes compared to the full data set.*

We are primarily interested in the interactions between the global transcriptional regulators, however both our PCA analysis and our clustering showed that the growth phase and biases from sample preparation had a strong effect on our gene expression patterns. To minimize the variance from these two sources, we reconstructed our data set after removing the first two principle components [147]. We then repeated our clustering analysis with this adjusted data set. Instead of finding 5 clusters as the optimal K value we found 7 clusters was optimal for both Hierarchical and K-means clustering (Figure 83). These clusters contained (25/934/286/83/429/483/1055) genes in each hierarchical cluster and (557/654/437/340/522/424/361) genes in each K-means cluster. When we generated a clustergram for our adjusted data set, we found that the large clusters of correlated genes were replaced with smaller groups of correlated genes (Figure 44). Additionally, while most of the exponential and stationary phase samples are still close in the clustergram, they are no longer clearly separated. It is perhaps unsurprising that many of the global regulators have different genetic programs for different growth states, given that fis and hns are expressed at different points during growth for example [148]. Additionally, the dendrogram has a much more asymmetrical pattern than with the full data set. Much of the regular patterning in the clustering of experiments was likely due to the strong impact of growth state and batch effects on the expression profiles, and the clustering of the effects of just the transcription factors is much more complex.

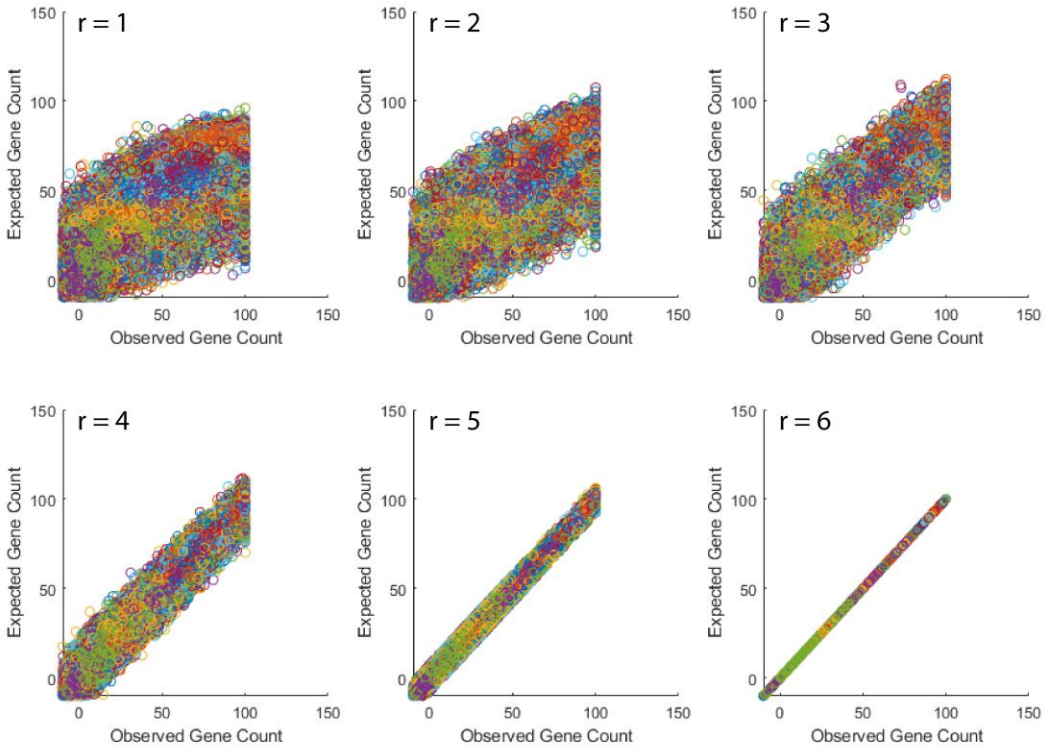



*Figure 45: Prediction of gene expression using r-th order epistasis coefficients after the first 2 principle components were removed from the data. Epistasis coefficients are determined by linear regression, where r = 6 calculates all possible epistasis coefficients and therefore perfectly matches the observed data. As r decreases, fewer coefficients are calculated. All gene counts are scaled to a range between 0-100 for comparison. The first two principle components were found to be associated with growth state ($OD_{595nm}$) and batch effects respectively. Their removal leads to a much more nuanced clustering of genes compared to the full data set. Despite the removal of these two components, the data shows the same characteristics as the full data set.*

We repeated our epistasis analysis using linear regression with this processed data set (Figure 45). The data showed the same epistasis pattern as will the full data set. This indicates that despite removing the component corresponding to growth state, we do not reduce the impact of the single coefficient corresponding to the combination of all perturbations and changing growth state (r = 5). Additionally, the variance in our data corresponding to growth state and batch effects do not seems to impact the higher order epistasis pattern in our data. The next step will be to individually remove coefficients independent from their order, to find the minimum number of coefficients to explain most of the data. By removing the smallest coefficients one at a time, we may find which higher order interactions are the most impactful.

Plotting the K-mean centroids showed much stronger patterns than the original data set Figure 46. We can see HNS perturbation patterns (where the oscillation is every other strain) in the first and third clusters. The FNR pattern in (every 2 strains) is in the seventh cluster. The Fis pattern (every 4 strains) is in the third and fifth clusters. The CRP pattern (every 8 strains) is in the fourth and sixth cluster. The ArcA pattern (every 16 strains) is in the first, second, fourth, and seventh clusters. We can also still see some responses that are conditional on the growth phase, despite the first principle component being removed.

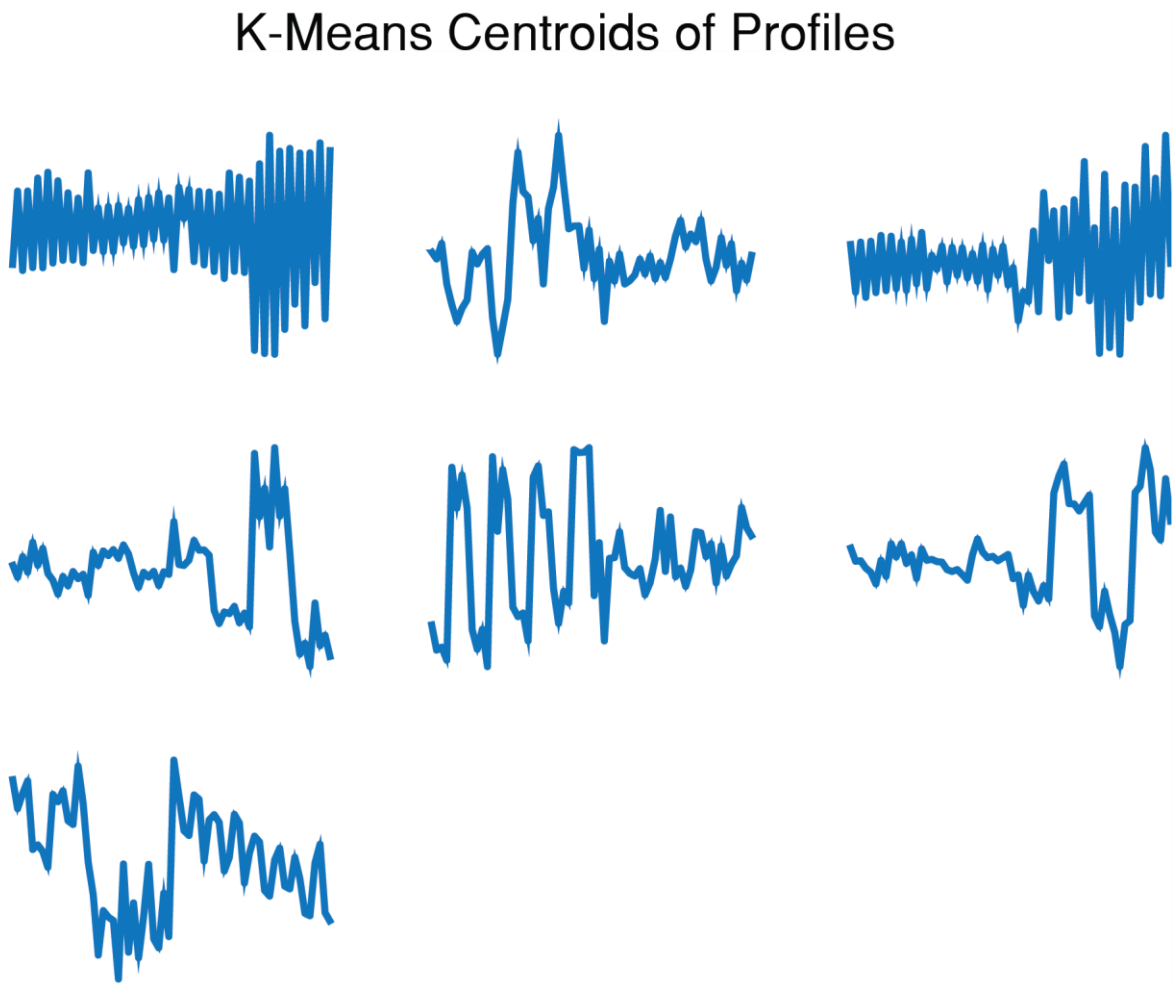



*Figure 46: K-mean centroids of pCKDL strains. Strains are ordered pCKDL 0 to 31 in exponential phase and then 0 to 31 in stationary phase. The pattern represents the mean expression for all genes within the cluster, with peaks indicating a higher expression and valleys representing lower expression.*

Our logic analysis of the genes clusters was also similar to the full data set, although here we found 7 main clusters of gene expression instead of 5. Generally, there are also more levels fitted to these clusters than with the full data set (Figure 47). While some of the plateaus seem to have simple logic (for example: the top plateau of the first cluster is !ArcA ∧ !HNS ∧ Stationary) however most still contain complex logic. We are currently working on a method to derive the disjunctive normal form from the truth tables that we have formed from the sorted graphs.

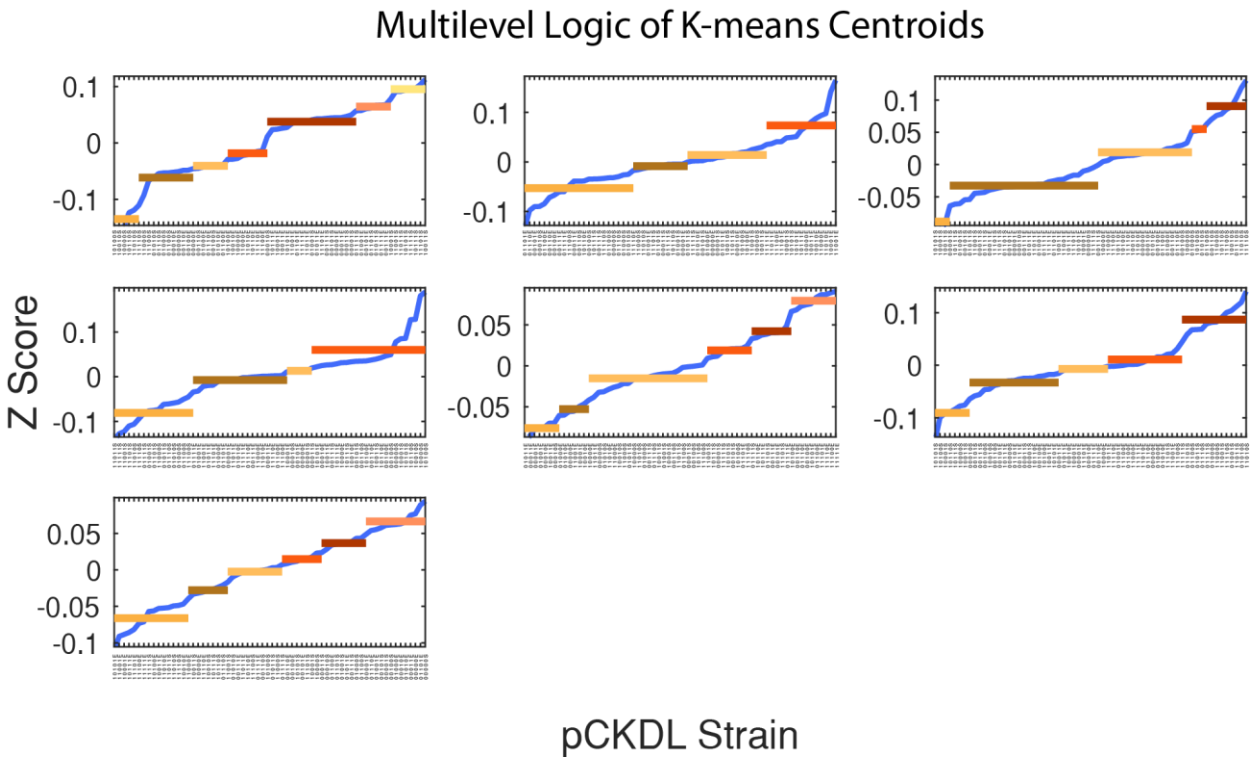



*Figure 47: Multi level logic of K-mean centroids after the first 2 principle components were removed from the data. Strains are sorted by their centroid. Plateaus are fitted by finding local maxima of the derivative of the curve, after smoothing. These local maxima are the locations of the steps. The perturbation code (x-axis) gives the logic which corresponds to each level of gene expression in each K-means cluster. The first two principle components were found to be associated with growth state ($OD_{595nm}$) and batch effects respectively. With these components removed, there are 2 more clusters identified, and typically clusters have more levels than the full data set.*

These results can already address one of our initial questions. How many dimensions $D_{out}$ are needed to represent most (ex: 90%) of the expression pattern of OUT genes? We observed that 22 dimensions were necessary to explain 75% of the variance in our data. Despite global regulators being coupled in strongly connected components within the transcriptional regulatory network, the majority of their impact on gene expression programs seems to be orthogonal. The coregulation of fis and crp for example, does not reduce the dimensionality of their response. In contrast, the interactions between global regulators increases the dimensionality of the response, rather than limiting it. This is because the epistatic interactions cause variance in the data that is not explained by their independent effects. Additionally, the expression of the global transcription factors is also a multi-level logic function. With different combinations of perturbations influencing the expression levels of the non-perturbed regulators. As a result, we still require a formalism to explain the complex gene expression programs we observe with a minimal system.

## 3.5 Logical Modelling of *E. coli* Regulatory Network

*With* ***Milan Lacassin***

Modeling of genetic regulatory networks often takes one of three forms: Thermodynamic, Differential equation-based, or Boolean models [58]. Thermodynamic models seek to explain how a gene will be activated or repressed, given a promoter and well-characterized transcription factors. It makes a key assumption that gene expression will be proportional to the number of transcriptional activators bound and inversely proportional to the number of repressors bound. This approach enables a detailed analysis of cis-element regulation, however it is highly dependent on functional binding sites and can have difficulty with context specific effects [58]. In contrast, Differential equation-based models generally do not consider the extreme detail thermodynamic approaches but in exchange are able to capture the dynamic nature of biological systems. They have successfully been used to model individual regulatory pathways such as regulation of the lack operon, however the large number of parameters makes computation prohibitive for larger systems containing hundreds of molecules [58]. This problem is addressed with Boolean or logical modeling [59]. These models represent regulatory interactions as logic gates with discrete states for each item. This simplification makes logical models easier to analyze and extend to large biological networks. This framework allows for the study of how a given state evolves through a network, which states of the network are stable, and the location of cyclic attractors. It can also be used to study perturbations, as nodes in the network can be fixed to a given state [58].

***Implementation of the E. coli transcriptional regulatory network in GINsim***

We used GINSim [59], with assistance from Denis Thieffry at ENS to make a logical model of the transcriptional network of *E. coli*. We focused on the dimensionality of the stable states space. This is done by comparing the dimensions of the stable state space for different perturbations to the unperturbed network. Combinations of perturbations can be characterized by the number of states they allow. We defined default logical rules for nodes of the network. Transcriptional inhibitors suppress activators, such that Inhibitors always set the node to 0, while activators will only set a node to 1 in the absence of any inhibitor. For a node to be turned on, it requires at least one activator and no repressors. In the absence of any activators or repressors, we tested two scenarios: the nodes are free (such that both states are possible) or the nodes are set to a basal value of 0. Finally, any complexes such at heterodimers or transcription factors requiring a co-factor, would only be functional in the presence of all required components. The logical functions are written as follows:

• Default function for complexes (without basal value):

$$N(t+1) = c_1(t) \wedge \cdots \wedge c_{n_c}(t) \wedge \left(N(t) \vee a_1(t) \vee \cdots \vee a_{n_a}(t)\right) \wedge !\, i_1(t) \wedge \cdots \wedge !\, i_{n_i}(t)$$

• Default function for complexes (with basal value):

$$N(t+1) = c_1(t) \wedge \cdots \wedge c_{n_c}(t) \wedge \left(a_1(t) \vee \cdots \vee a_{n_a}(t)\right) \wedge !\, i_1(t) \wedge \cdots \wedge !\, i_{n_i}(t)$$

• Default function for normal nodes (without basal value):

$$N(t+1) = \left(N(t) \vee a_1(t) \vee \cdots \vee a_{n_a}(t)\right) \wedge !\, i_1(t) \wedge \cdots \wedge !\, i_{n_i}(t)$$

• Default function for normal nodes (with basal value):

$$N(t+1) = \left(a_1(t) \vee \cdots \vee a_{n_a}(t)\right) \wedge !\, i_1(t) \wedge \cdots \wedge !\, i_{n_i}(t)$$

Here ❑∧❑ denotes the logical operator AND, ❑∨❑ denotes the logical operator OR, ! denotes the logical operator NOT, $N(t+1)$ denotes the state of node N at time t+1, $N(t)$ denotes the state of node N at time t, $c_n(t)$ is the state of the n-th component of the complex N at time t, $a_n(t)$ is the state of the n-th activator of node N at time t, $i_n(t)$ is the state of the n-th inhibitor of node N at time t, $n_c$ is the number of components of the complex N, $n_a$ is the number of activators of the node N, and $n_i$ is the number of inhibitors of the node N.

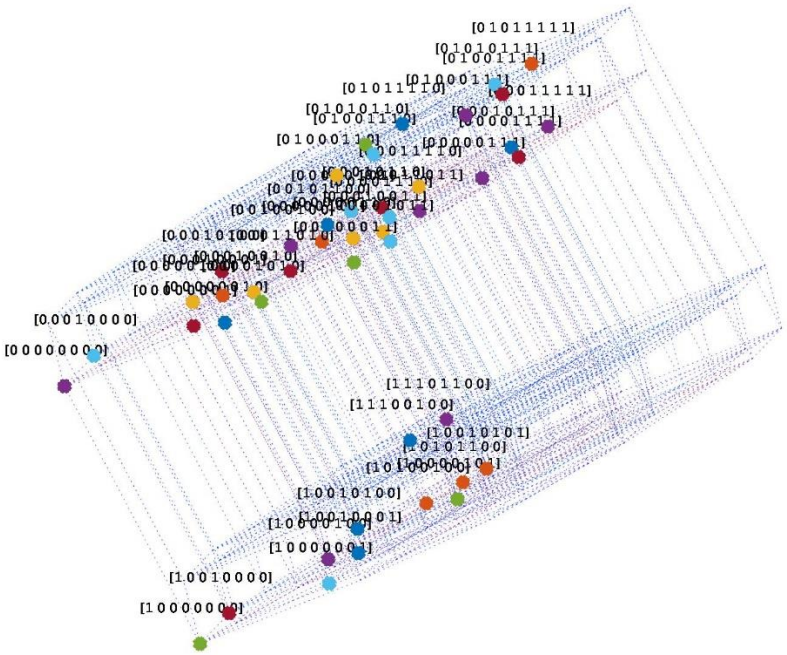



*Figure 48: Hypercube representing all reachable stable states of the energy cluster in E. coli transcriptional regulatory network. Possible steady states are highlighted, with the Boolean state indicated. Each bit corresponds to one transcription factor or small regulatory molecule such as ppGpp or cAMP.*

Due to the hierarchy of the strongly connected components in the transcriptional regulatory network of *E. coli*, we are able to model the upstream cluster first, and use the outputs as inputs for modeling the downstream clusters. This allows us to assemble the stable states of the complete system by taking all combinations of the stable states where common nodes have the same values. When we consider N genes of interest, the state space exists within a hypercube of N dimensions. Due to the smaller size of several clusters this can be represented on a hypercube projection (Figure 48), allowing us to visualize the impact of perturbations and their dimensionality compared to the reference system. We are also able to compute the number of states forbidden, and the number of new states allowed by each perturbation. We found that perturbing the genes within a strongly connected component increased the dimensionality of the possible state spaces. It is possible for us to expand this model into multi-level logic using the transcriptional data we have obtained, however this has not yet been implemented.

## 3.6 Methodology

*Creation of Vectors*

Lun Cui and David Bikard provided us with pCRRNA. We removed the upstream promoter region of pCRRNA by PCR with primers from the crRNA spacer to the origin of replication, excluding the sequence we wished to remove. A bidirectional terminator and 3' priming sequence was synthesized as a gBlock and added to the PCR fragment using Gibson assembly. This resulted in pCKDL, which contains a single CRISPR repeat and a dual BsaI cloning site. Spacers were synthesized as gBlocks, which also contained a CRISPR repeat, another cloning site (either BsaI or BsmBI), and a DNA barcode. These gBlocks were flanked with either BsaI cut sites or BsmBI sites (the opposite of their cloning site). Each gBlock was cloned into pJET1.2 cloning vector with blunt ligation and sequenced before use. They were then extracted from pJET1.2 with PCR, and cloned into pCDKL using golden gate assembly. This mix contained 1 µL of fast digest buffer, 1 µL of ATP (10 mM), 1 µL of either Eco31I (BsaI) or Esp3I (BsmBI), 1 µL of T4 Ligase (30 Weiss U), 3 µL of pCKDL and 3 µL of PCR insert. Golden gate solution was incubated for 25 cycles of 4 minutes at 37 $^{o}$C and 3 minutes at 16 $^{o}$C, before heat inactivation. Clones were transformed into chemically competent Oneshot Top10 *E. coli* cells. Plasmids were extracted using NucleoSpin Plasmid Miniprep kits from Macherey-Nagel and sequenced by GATC-biotech.

*Growth Conditions of Cultures*

The host strain for all pCKDL strains is LC-E24 :: dcas9 2tetO HK022 attB, which was provided by Lun Cui and David Bikard at Institute Pasteur. This strain is derived from MG1655 and has dcas9 under a tetracycline inducible promoter (derived from pdCas9 with an extra tetO site) integrated into the genome. Glyercol stocks of each culture were streaked onto individual Lysogeny Broth (LB) agar plates containing 50 µg/mL of Kanamycin. Single colonies were inoculated into 2 mL of selected media (either LB or M9 supplemented with 0.4% Glucose or Lactose) containing 50 µg/mL Kanamycin. Cultures were placed in a 37$^{o}$C incubator for either overnight for 16 hours for LB cultures or for 24 hours for M9 cultures.

*Quantification of Growth Fitness*

Assays were performed by filling each well of a Greiner, 96 Well, F-Bottom, clear microplate with 198 µL of growth media supplemented with 50 µg/mL Kanamycin and 250 ng/mL anhydrotetracycline. For each pCKDL knockdown strain, 2 µL of pre-culture was inoculated into 3 individual wells. A volume of 50 µL of mineral oil was added to each well of the microplate. The Absorbance at 595 nm was recorded every 10 minutes for 20 hours with a SpectraMax i3x. Microplates were incubated at 37$^{o}$C and shook for 300 seconds after each measurement. To determine the fitness measurement of each knockdown, the maximum derivative of the log2 of the OD was determined to locate the middle of exponential growth (omitting the first 5 measurements to reduce noise from measurements taken below the detection threshold). The slope of the growth curve was then fit to $\pm$ 20 minutes of this point to determine the Exponential growth rate fitness. The maximum OD after 20H of growth was taken as the second fitness measurement. Fitness measurements for each knockdown were made for LB media, and M9 media containing 0.4% of either glucose or lactose.

*Transcription quantification with RNAtag-Seq*

Cultures were made by diluting 20 µL of pre culture into 2 mL of selected media (LB or M9 with 0.4% Glucose or Lactose). Cultures were then grown to mid-exponential phase and early stationary phase. One milliliter of culture was pipetted into a microcentrifuge tube and centrifuged for 1 minute at 8,000 RCF. Five hundred microliters of supernatant was removed from each tube and the cells were resuspended. One milliliter of RNAprotect Bacteria from Qiagen was added to each tube, mixed well by inversion, and incubated at room temperature for 5 minutes. Tubes were then centrifuged at 8000 RCF for 10 minutes. The supernatant was removed and replaced with 100 µL of TE buffer with 15 mg/mL of Lysozyme and 10 µL of Proteinase K. Cultures were incubated at room temperature for 10 minutes, after which 350 µL of RLT Buffer with β-Mercaptoethanol was added. Each tube was vortexed before adding 250 µL of Ethanol. Each tube was mixed by inversion and the solution was loaded onto an RNeasy Column. Columns were centrifuged for 15 seconds at 8000 RCF and the flow through was discarded. For each column, 700 µL of Buffer RW1 was added, and then centrifuged for 15 seconds at 8000 RCF and the flow through was discarded. Columns were then washed twice by adding 500 µL of Buffer RPE and centrifuging for 15 seconds at 8000 RCF before discarding the flow through. Columns were spun for an additional 1 minute to dry the membrane before discarding the collection tube and putting the column into a new 1.5 mL micro centrifuge tube. RNA was eluted by adding 30 µL of Nuclease Free water onto the membrane and spinning for 1 minute at 8000 RCF. RNA yield and quality were quantified with a NanoDrop and Agilent Tapestation 4200; 600 ng of RNA was aliquoted into a tube before increasing the volume to 15 µL with Nuclease free water and the RNA integrity number was recorded for each sample. For each tube, 1 µL of SUPERase-IN was added and samples were frozen overnight at -80$^{o}$C. The next day, samples were thawed and 4 µL of FastAP buffer was added. Samples were incubated on pre-heated thermal cycler for 3 minutes at 92$^{o}$C to fragment the RNA. DNase and FastAP treatment was then performed by adding 1 µL of RNase Inhibitor, Murine, 4 µL of Turbo DNase, 10 µL of FastAP, and 5 µL of Nuclease free water to each sample, mixing and incubating for 30 minutes at 37$^{o}$C. Samples were then cleaned by adding 40 µL of Nuclease free water and 160 µL of Agencourt RNAClean XP beads. Samples were incubated at room temperature for 15 minutes to allow the RNA to bind to the beads before placing on a magnet for 5 minutes. The solution was then removed and replaced with 200 µL of fresh 70% EtOH. The wash was then removed and replaced with another 200 µL of fresh 70% EtOH. The second wash was removed and the beads were allowed to air dry for 10 minutes. 12 µL of Nuclease free water was added to the beads and they were removed from the magnet. Random samples were checked for their fragmentation profile in each batch on the Agilent Tapestation 4200; 5 µL of each sample was carried forward to the adapter ligation while 1 µL of SUPERase-IN was added to the remaining sample and it was frozen at -80$^{o}$C. For adapter ligation, 1 µL of barcoded adapater was added to each 5 µL sample of RNA and heated to 70$^{o}$C for 2 minutes before being placed back onto ice. Ligation mix was made by mixing 80 µL of 10x T4 RNA Ligase Buffer, 72 µL DMSO, 8 µL ATP, 320 µL PEG 8000, 12 µL RNAse inhibitor, Murine, and 72 µL T4 RNA Ligase 1. For each sample, 14.1 µL of ligation mix was added and mixed very well. Each sample was incubated at 22$^{o}$C for 1.5 hours. Samples were then pooled by adding 60 µL of RLT buffer and 160 µL of 1:1 binding buffer:EtOH to each sample and mixed in a 5 mL Eppendorf tube. Samples were loaded onto Zymo Clean & Concentrator columns with a vacuum manifold. RNA was eluted by adding 14 µL of Nuclease free water twice for a total volume of 28 µL. Ribosomal RNA was removed using Ribo-Zero Magnetic Kit (Bacteria) from Illumina. Magnetic beads were prepared by adding 225 µL of Magnetic beads to micro centrifuge tube and placing on a magnetic rack for 1 minute. Beads were then washed twice with 225 µL of Nuclease

free water before being resuspended in 65 µL of Resuspension Solution and 1 µL of RiboGuard RNase Inhibitor. Pooled RNA was treated by mixing 26 µL of RNA solution with 10 µL of rRNA Removal Solution and 4 µL of Reaction Buffer, before incubating at 65$^{o}$C for 10 minutes and then 5 minutes at room temperature. Washed Magnetic beads were then added to the RNA solution, vortexed, and incubated at room temperature for 5 minutes before increasing the temperature to 50$^{o}$C and incubating for a further 5 minutes. Samples were then placed on a magnetic rack and the supernatant was transferred to a new RNase-free tube. The rRNA free supernatant was then cleaned using AMPure XP beads by adding 160 µL of beads to the RNA solution. The Solution was incubated for 15 minutes at room temperature and placed on a magnetic rack for 5 minutes. The supernatant was removed and replaced with 200 µL of fresh 80% EtOH. The EtOH was removed and replaced with another 200 µL of 80% EtOH. All EtOH was removed and the beads were allowed to air dry. Beads were removed from the magnet and 14 µL of Nuclease free water was added and mixed well to Elute RNA. The solution was placed back on the magnetic rack and supernatant containing the RNA was removed and transferred to a fresh tube. First strand cDNA synthesis was performed by adding 2 µL of AR2 Primer (oMD667) to the RNA sample, mixing, and heating to 70$^{o}$C for 2 minutes before placing immediately back onto ice. The RT mix was made by adding 2 µL 10x Affinity Script RT Buffer, 2 µL DTT, 0.8 µL dNTP mix, 0.4 µL RNase inhibitor, murine, and 0.8 µL AffinityScript RT Enzyme to the RNA sample. Samples were mixed well and quickly centrifuged for 5 seconds before being placed into a preheated thermocycler at 55$^{o}$C for 55 minutes. RNA was degraded from cDNA by adding 2 µL of fresh 1M NaOH to each sample and incubating at 70$^{o}$C for 12 minutes. Samples were neutralized with 4 µL of 0.5M Acetic Acid. Sample volume was increased to 40 µL by adding 14 µL of Nuclease free water and transferring to a new tube. Samples were cleaned by adding 80 µL of RNAClean XP beads to each sample, mixing, and incubating at room temperature for 15 minutes. Samples were placed on a magnetic rack for 5 minutes and supernatant was discarded. Beads were then washed with 200 µL of fresh 70% EtOH. The EtOH was removed and replaced with another 200 µL of 70% EtOH. All EtOH was removed and the beads were allowed to air dry for 10 minutes. The second adaptor ligation was done by adding 5 µL of Nuclease free water to the beads to elute the DNA, and then adding 2 µL of 3Tr3 Adaptor (oMD668) to the cDNA and magnetic bead solution. The solution was then heated for 3 minutes at 75$^{o}$C and then mixed before adding ligation mix consisting of 2 µL of 10x T4 Ligase Buffer, 0.8 µL DMSO, 0.2 µL ATP, 8.5 µL PEG8000, and 1.5 µL T4 RNA Ligase 1. The Solution was mixed well and incubated overnight at 22$^{o}$C. Ligations were cleaned by adding 80 µL of fresh RNAClean XP beads to each sample, mixing, and incubating at room temperature for 15 minutes. Samples were placed on a magnetic rack for 5 minutes and supernatant was discarded. Beads were then washed with 200 µL of fresh 70% EtOH. The EtOH was removed and replaced with another 200 µL of 70% EtOH. All EtOH was removed and the beads were allowed to air dry for 10 minutes. DNA was eluted by adding 25 µL of Nuclease free water, mixing well, and placing sample on a magnetic rack. The supernatant containing the cDNA was then transferred to a new tube and cleaned again by adding 37.5 µL of fresh RNAClean XP beads to each sample, mixing, and incubating at room temperature for 15 minutes. Samples were placed on a magnetic rack for 5 minutes and supernatant was discarded. Beads were then washed with 200 µL of fresh 70% EtOH. The EtOH was removed and replaced with another 200 µL of 70% EtOH. All EtOH was removed and the beads were allowed to air dry for 3 minutes. DNA was eluted by adding 25 µL of Nuclease free water, mixing well, and placing sample on a magnetic rack; the supernatant containing the cDNA was then transferred to a new tube. The number of PCR cycles needed to enrich the sample sufficiently for sequencing was determined by Phusion PCR with 2P_univP5 (oMD669) and 2P_bacrode (oMD700) for 9, 12, and 15 cycles. PCRs were

performed with a Tm of 55$^{o}$C and an extension time of 60 seconds. PCRs were purified by increasing the volume to 25 µL with Nuclease free water, adding 37.5 µL AMPure beads, mixing, and incubating at room temperature for 15 minutes. The PCR solution was then placed on a magnetic rack for 5 minutes and the supernatant removed. The beads were washed with 200 µL of fresh 70% EtOH twice and then allowed to air dry for 10 minutes before being eluted with 10 µL of low TE (10 mM Tris, 0.1M EDTA). PCR products were then quantified on the Agilent TapeStation 4200 to determine the minimum required number of PCR cycles. PCR was then performed on 10 µL of cDNA using 2P_univP5 (oMD669, oMD687-oMD693) and 2P_bacrode (oMD670, oMD672-oMD686) with Phusion Master Mix with the determined minimum number of cycles, a Tm of 55$^{o}$C, and an extension time of 60 seconds. PCRs were purified by adding 75 µL AMPure beads, mixing, and incubating at room temperature for 15 minutes. The PCR solution was then placed on a magnetic rack for 5 minutes and the supernatant removed. The beads were washed with 200 µL of fresh 70% EtOH twice and then allowed to air dry for 10 minutes before being eluted with 25 µL of Nuclease free water. DNA was cleaned again by adding 17.5 µL of AMPure beads, mixing, and incubating at room temperature for 15 minutes. The PCR solution was then placed on a magnetic rack for 5 minutes and the supernatant removed. The beads were washed with 200 µL of fresh 70% EtOH twice and then allowed to air dry for 10 minutes before being eluted with 10 µL of low TE (10 mM Tris, 0.1M EDTA). DNA was then quantified on the Agilent TapeStation 4200, and Qubit 3 Fluorometer, and diluted to 10 nM concentration.

## 3.7 Supplementary Figures

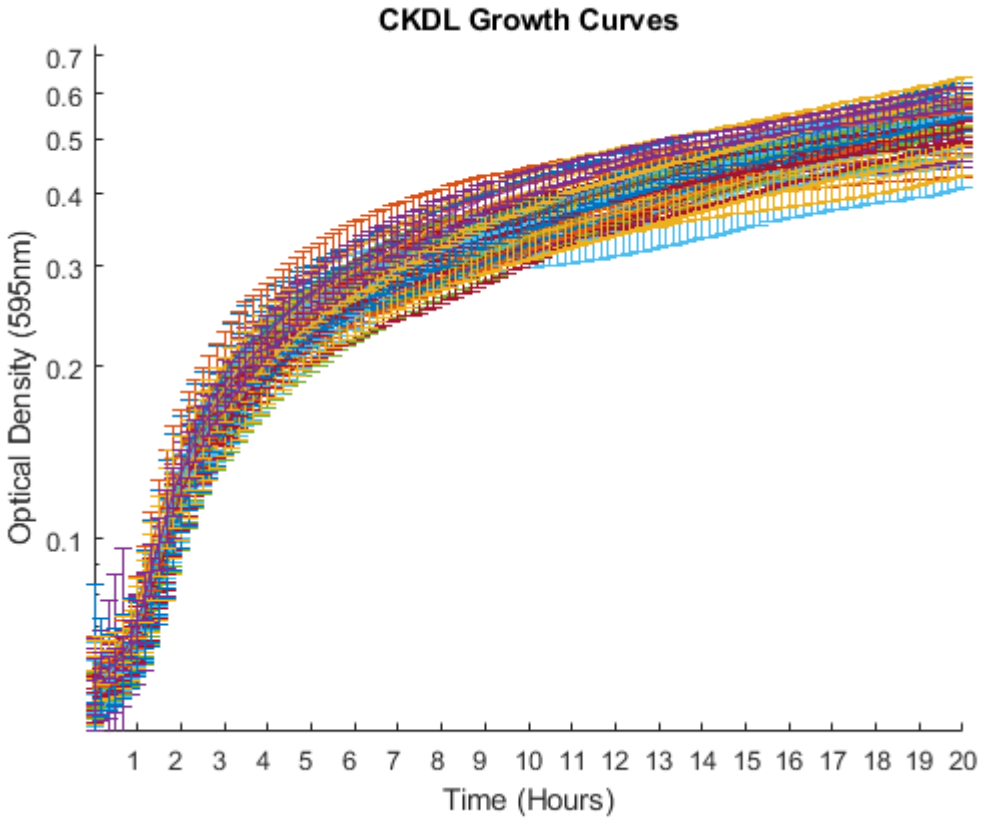



*Figure 49: Growth Curves for all 32 pCKDL strains grown in LB. E. coli strains bearing pCKDL vectors were grown 20 hours in 200 µL of LB media while recording the optical density. There are three biological replicates taken from individual single colonies each with their own three technical replicates for a total of nine. The median OD is plotted with error bars representing the standard deviation.*

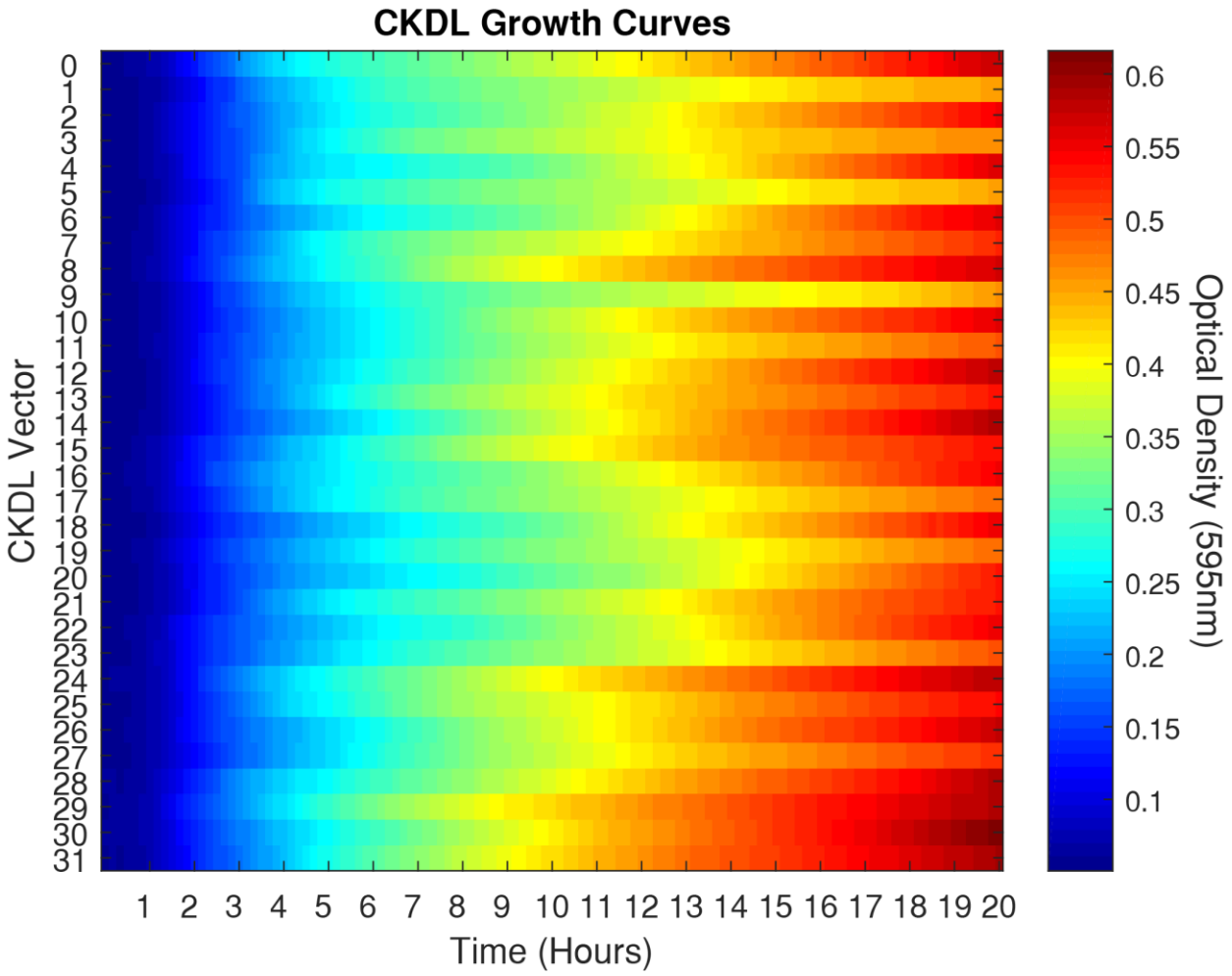



*Figure 50: Growth Curves for all 32 pCKDL strains in LB. Rows represent different pCKDL strains median OD over 20 hours of growth.*

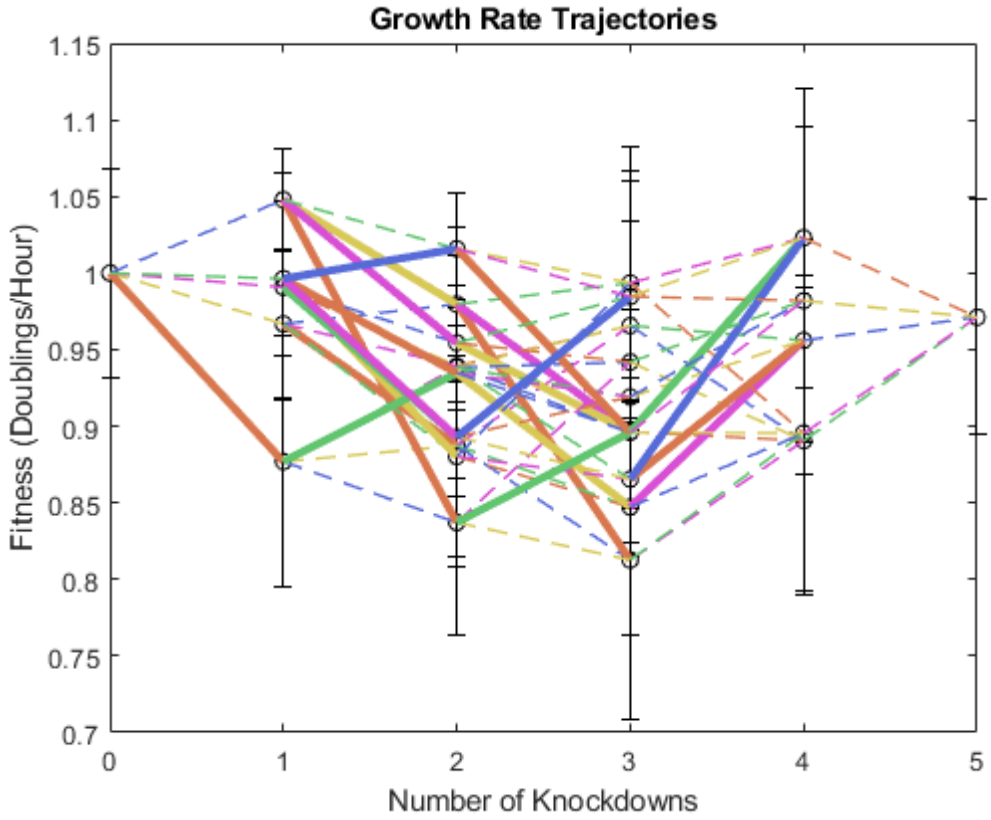



*Figure 51: Fitness of exponential growth rate for pCKDL in LB media by number of Knockdowns. pCKDL strains are arranged by the number targeting spacers in the CRISPR array. Circles indicate the median fitness measurement of a pCKDL strain, normalized by dividing the fitness of the pseudo-wild-type pCKDL 0. Error bars represent standard deviation. Solid lines are significantly ($p < 0.05$) different according to Welch's t-test. Lines are colored by the targeting spacer which differs between the two pCKDL strains (green: arcA, blue: crp, pink: fis, orange: fnr, yellow: hns). Data represents 3 biological replicates for each pCKDL strain and three technical replicates for each biological replicate (n = 9), error bars represent standard error of the median.*

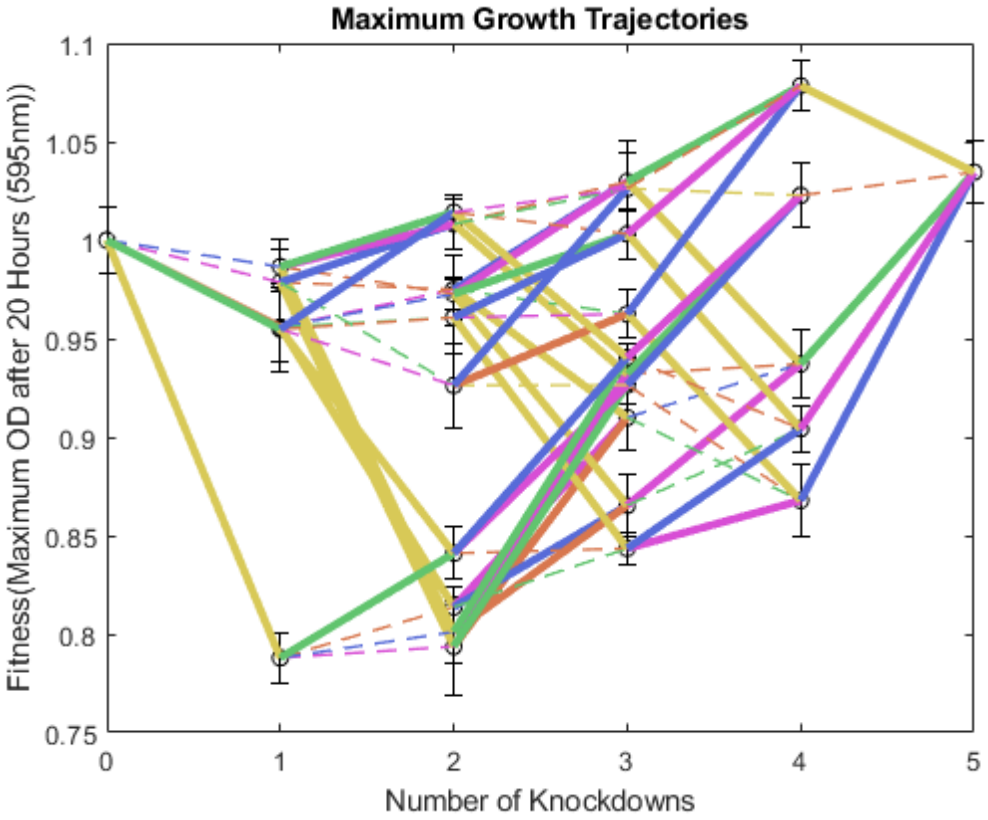



*Figure 52: Fitness of Maximum OD reached after 20 hours for pCKDL in LB media by number of Knockdowns. pCKDL strains are arranged by the number targeting spacers in the CRISPR array. Circles indicate the median fitness measurement of a pCKDL strain, normalized by dividing the fitness of the pseudo-wild-type pCKDL 0. Error bars represent standard deviation. Solid lines are significantly ($p < 0.05$) different according to Welch's t-test. Lines are colored by the targeting spacer which differs between the two pCKDL strains (green: arcA, blue: crp, pink: fis, orange: fnr, yellow: hns). Data represents 3 biological replicates for each pCKDL strain and three technical replicates for each biological replicate (n = 9), error bars represent standard error of the median.*

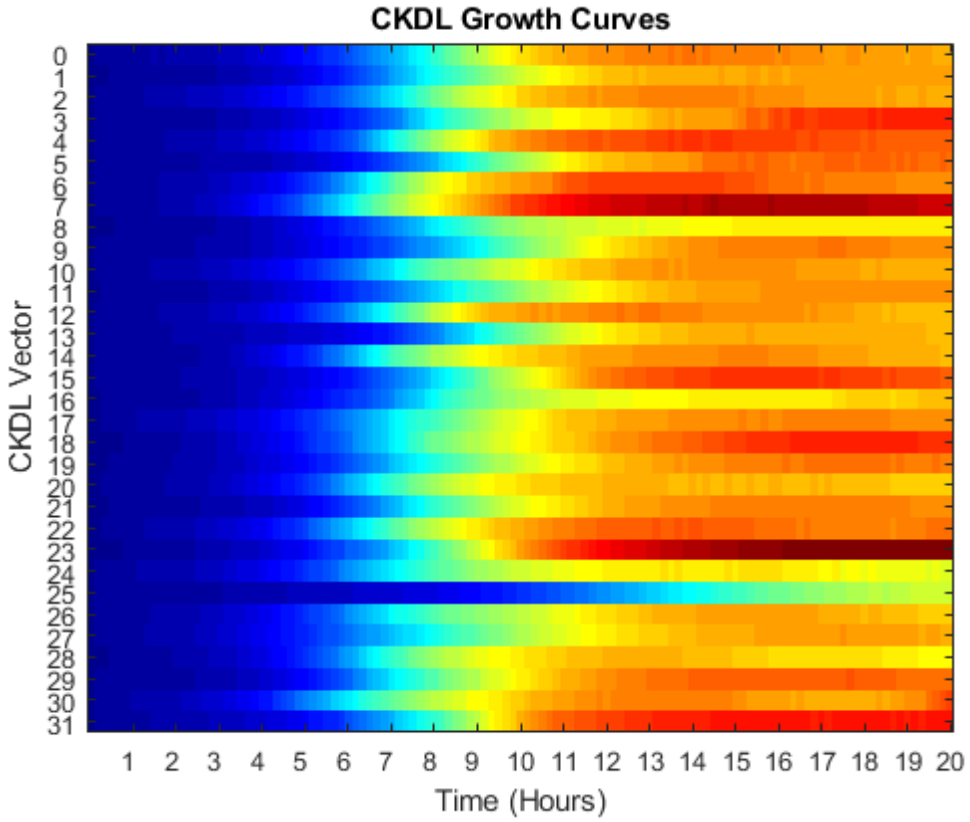



*Figure 53: Growth Curves for all 32 pCKDL strains in M9 media supplemented with 0.4% glucose. Rows represent different pCKDL strains median OD over 20 hours of growth.*

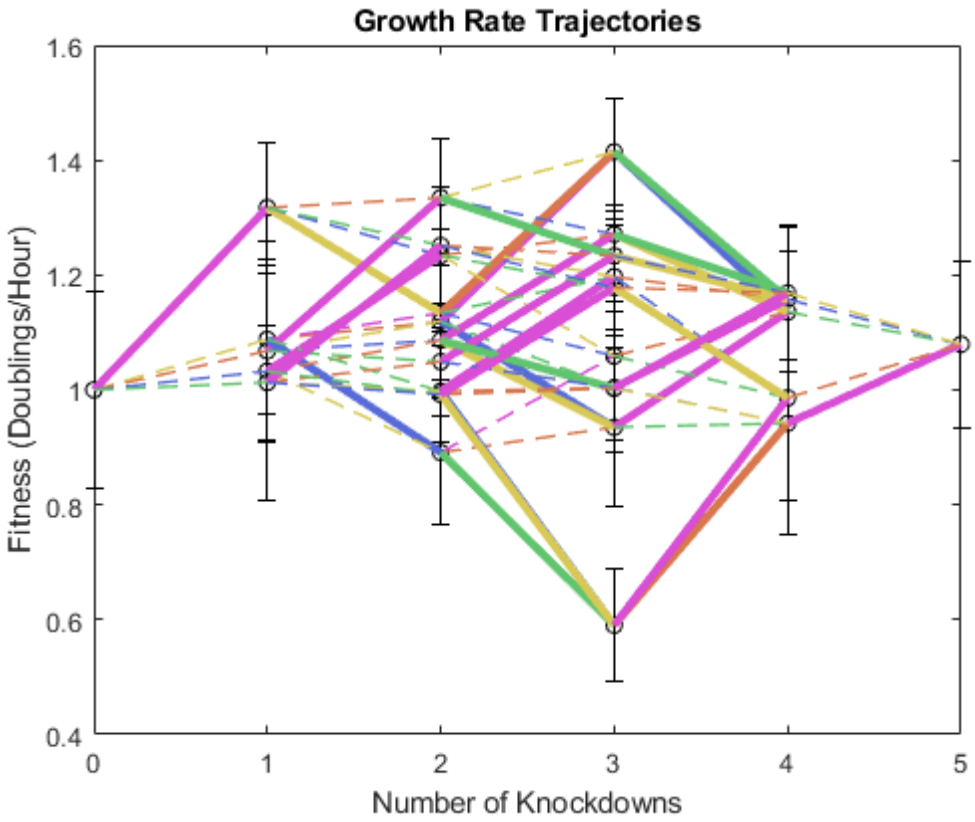



*Figure 54: Fitness of exponential growth rate for pCKDL in M9 media supplemented with 0.4% glucose by number of Knockdowns. pCKDL strains are arranged by the number targeting spacers in the CRISPR array. Circles indicate the median fitness measurement of a pCKDL strain, normalized by dividing the fitness of the pseudo-wild-type pCKDL 0. Error bars represent standard deviation. Solid lines are significantly ($p < 0.05$) different according to Welch's t-test. Lines are colored by the targeting spacer which differs between the two pCKDL strains (green: arcA, blue: crp, pink: fis, orange: fnr, yellow: hns). Data represents 3 biological replicates for each pCKDL strain and three technical replicates for each biological replicate (n = 9).*

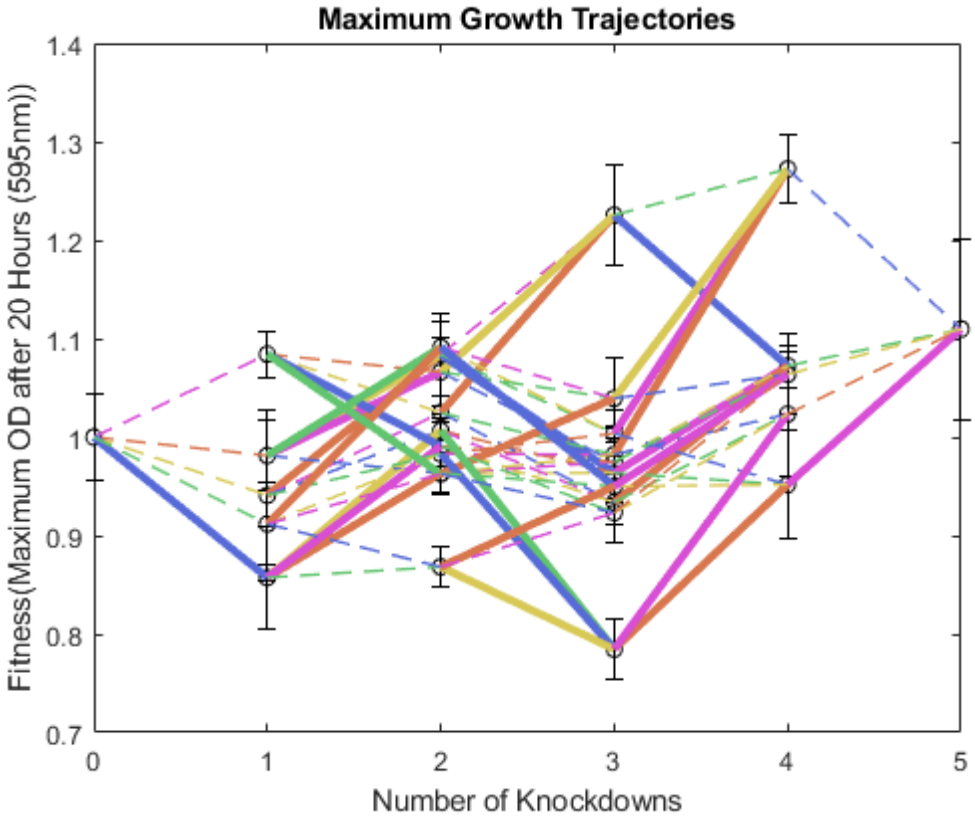



*Figure 55: Fitness of Maximum OD reached after 20 hours for pCKDL in M9 media supplemented with 0.4% glucose by number of Knockdowns. pCKDL strains are arranged by the number targeting spacers in the CRISPR array. Circles indicate the median fitness measurement of a pCKDL strain, normalized by dividing the fitness of the pseudo-wild-type pCKDL 0. Error bars represent standard deviation. Solid lines are significantly ($p < 0.05$) different according to Welch's t-test. Lines are colored by the targeting spacer which differs between the two pCKDL strains (green: arcA, blue: crp, pink: fis, orange: fnr, yellow: hns). Data represents 3 biological replicates for each pCKDL strain and three technical replicates for each biological replicate (n = 9).*

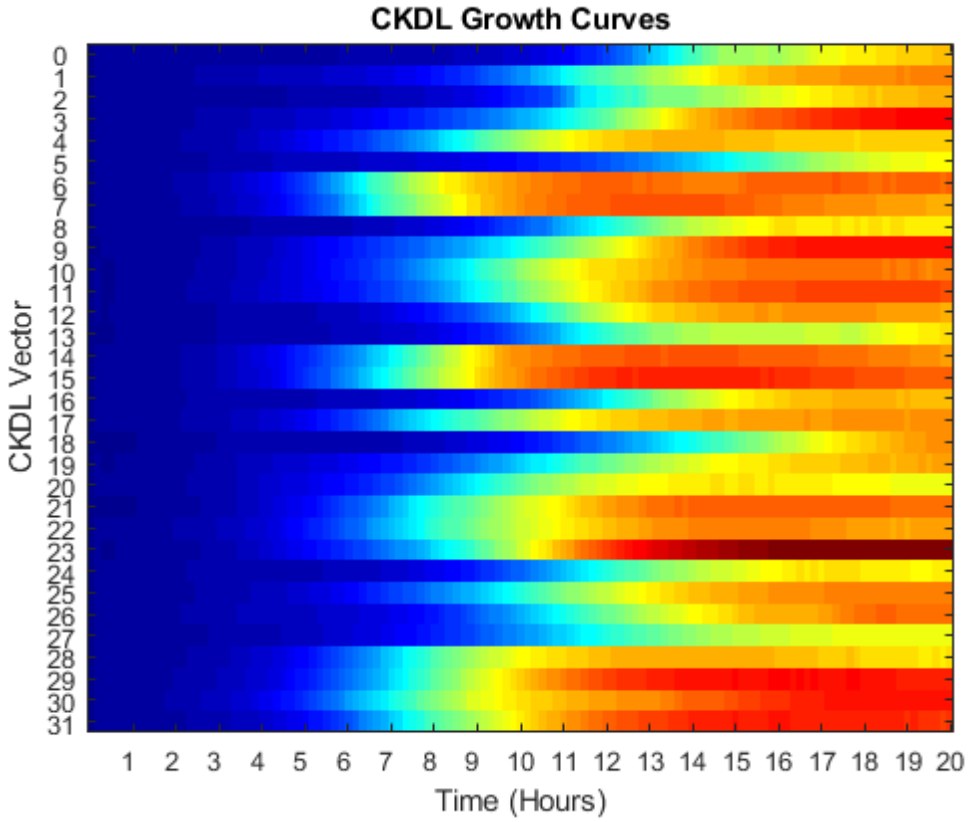



*Figure 56: Growth Curves for all 32 pCKDL strains in M9 media supplemented with 0.4% lactose. Rows represent different pCKDL strains median OD over 20 hours of growth.*

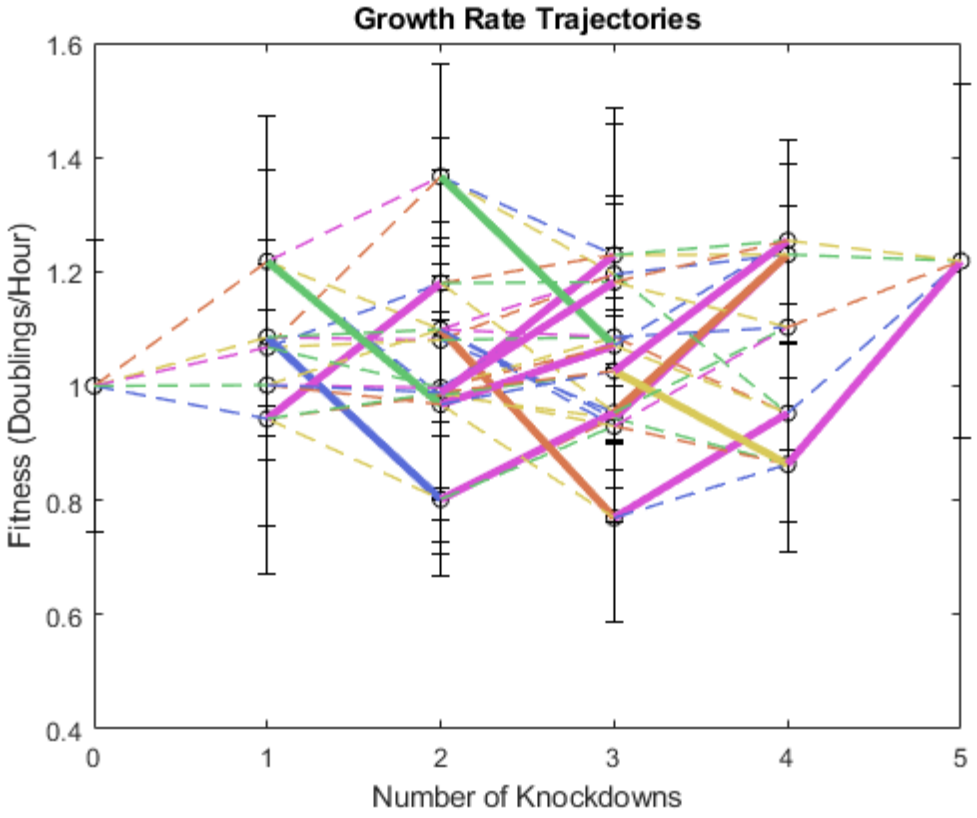



*Figure 57: Fitness of exponential growth rate for pCKDL in M9 media supplemented with 0.4% lactose by number of Knockdowns. pCKDL strains are arranged by the number targeting spacers in the CRISPR array. Circles indicate the median fitness measurement of a pCKDL strain, normalized by dividing the fitness of the pseudo-wild-type pCKDL 0. Error bars represent standard deviation. Solid lines are significantly ($p < 0.05$) different according to Welch's t-test. Lines are colored by the targeting spacer which differs between the two pCKDL strains (green: arcA, blue: crp, pink: fis, orange: fnr, yellow: hns). Data represents 3 biological replicates for each pCKDL strain and three technical replicates for each biological replicate (n = 9).*

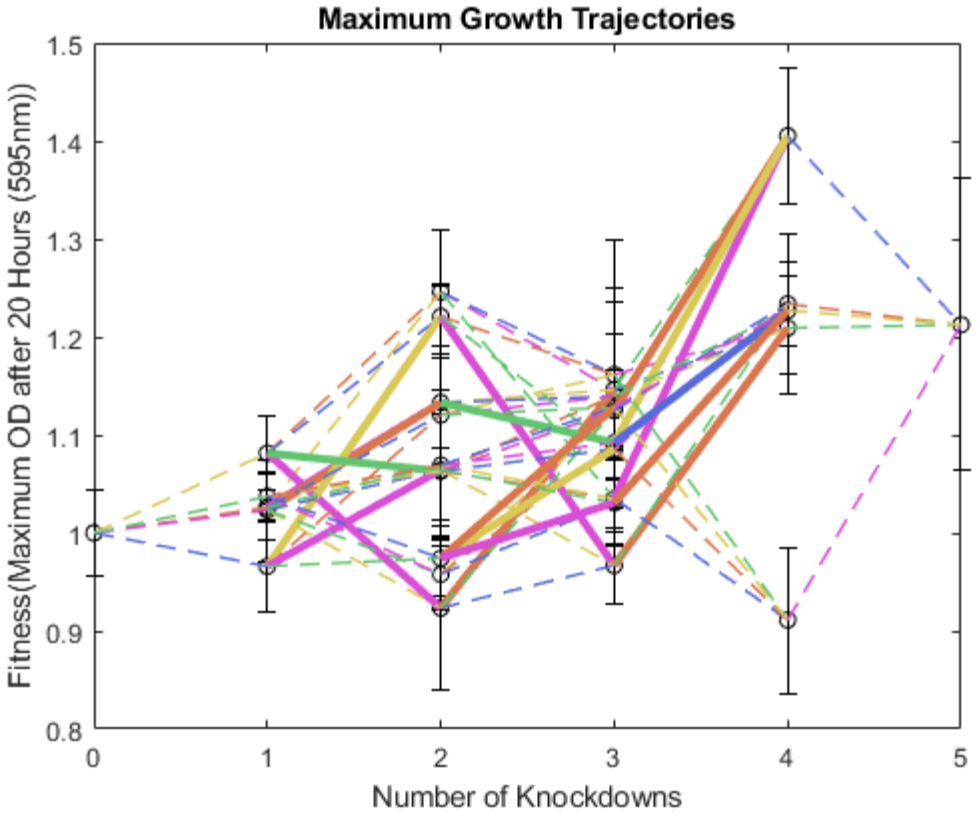



*Figure 58 Fitness of Maximum OD reached after 20 hours for pCKDL in M9 media supplemented with 0.4% lactose by number of Knockdowns. pCKDL strains are arranged by the number targeting spacers in the CRISPR array. Circles indicate the median fitness measurement of a pCKDL strain, normalized by dividing the fitness of the pseudo-wild-type pCKDL 0. Error bars represent standard deviation. Solid lines are significantly ($p < 0.05$) different according to Welch's t-test. Lines are colored by the targeting spacer which differs between the two pCKDL strains (green: arcA, blue: crp, pink: fis, orange: fnr, yellow: hns). Data represents 3 biological replicates for each pCKDL strain and three technical replicates for each biological replicate (n = 9).*

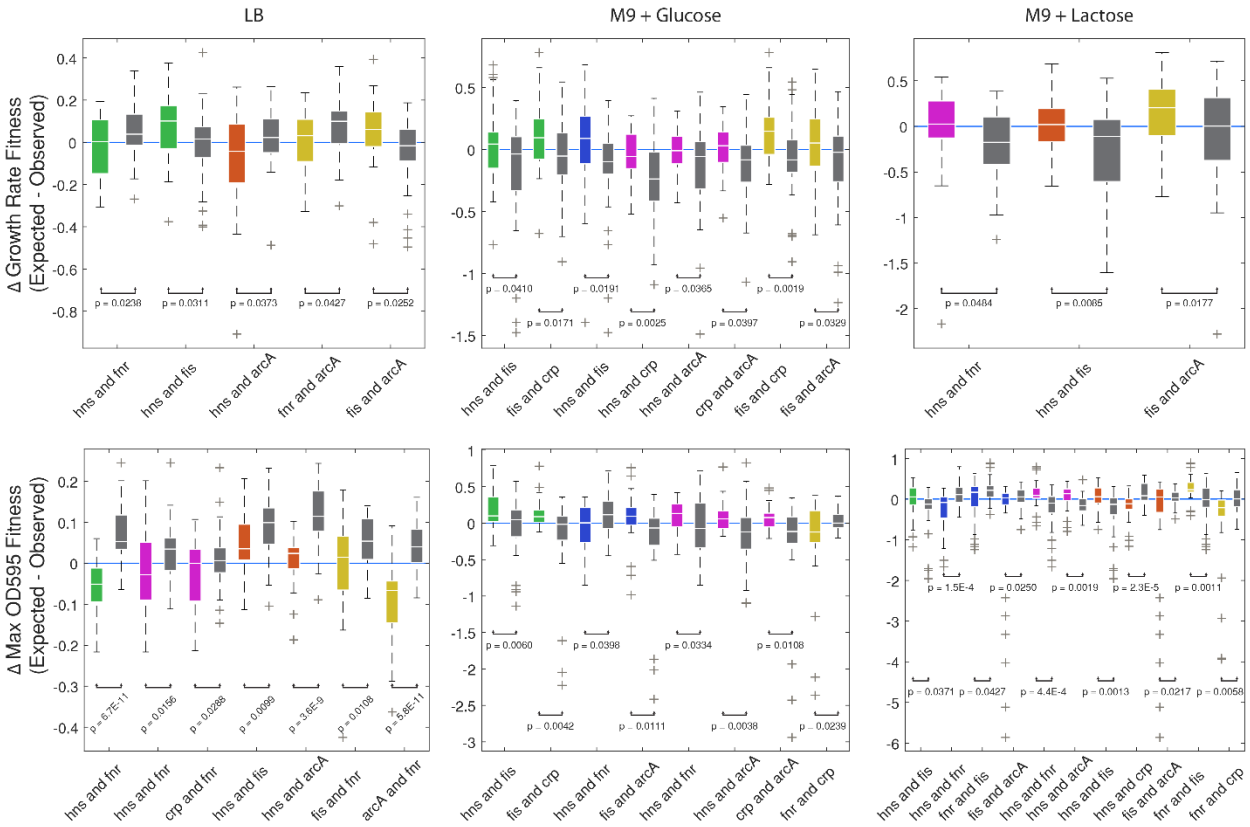



*Figure 59: Second order epistatic interactions in observed fitness phenotypes. For each pair of genes, the normalized fitness of the pCKDL strain with both genes perturbed is expected to be the product of the normalized fitness of the single perturbations. Each pair of genetic perturbations can be found in 8 backgrounds containing all possible combinations of the remaining 3 perturbations. The difference between the expected fitness and the observed fitness (y-axis) for the double perturbation (x-axis) is plotted for all 8 backgrounds, split between backgrounds with a third targeting spacer (coloured boxes; green: arcA, blue: crp, pink: fis, orange: fnr, yellow: hns) and a non-targeting spacer (dark grey boxes). Only interactions with a significant difference between the backgrounds with or without a third genetic perturbation are shown ($p < 0.05$) for a total of 42/180 possible conditions. Boxes represent the 25$^{th}$ and 75$^{th}$ percentile, the white line representing the median of the data, and whiskers represent the limits of the data not considered outliers. Outliers are plotted individually as grey '+'. The blue line indicates no epistasis.*

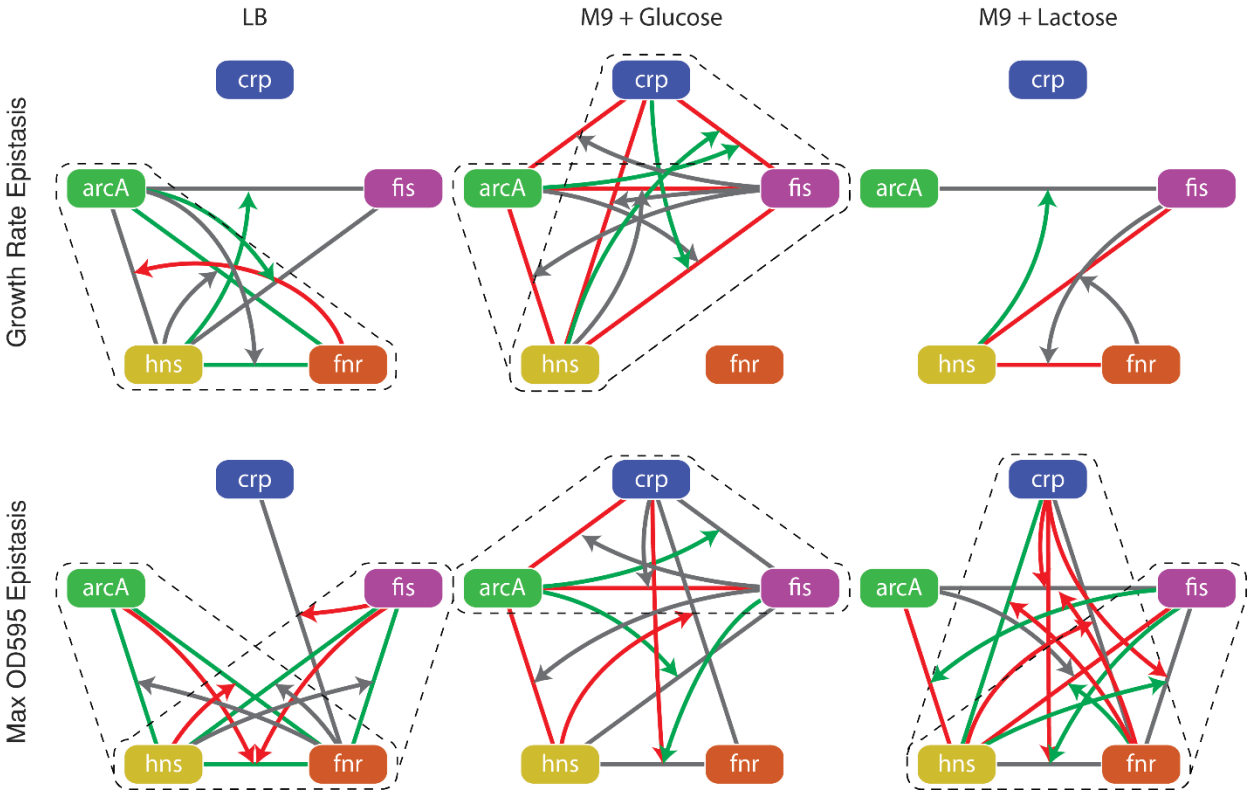



*Figure 60: Second order epistatic interactions between global transcription factors. Straight lines between coloured boxes represent the first order epistasis observed between two transcription factors. Grey indicates no epistasis, green indicates positive epistasis and red indicates negative epistasis. Curved arrows indicate when the presence of a transcription factor changes the epistasis of between two other transcription factors, with the color of the arrow reflecting the new epistasis status.*

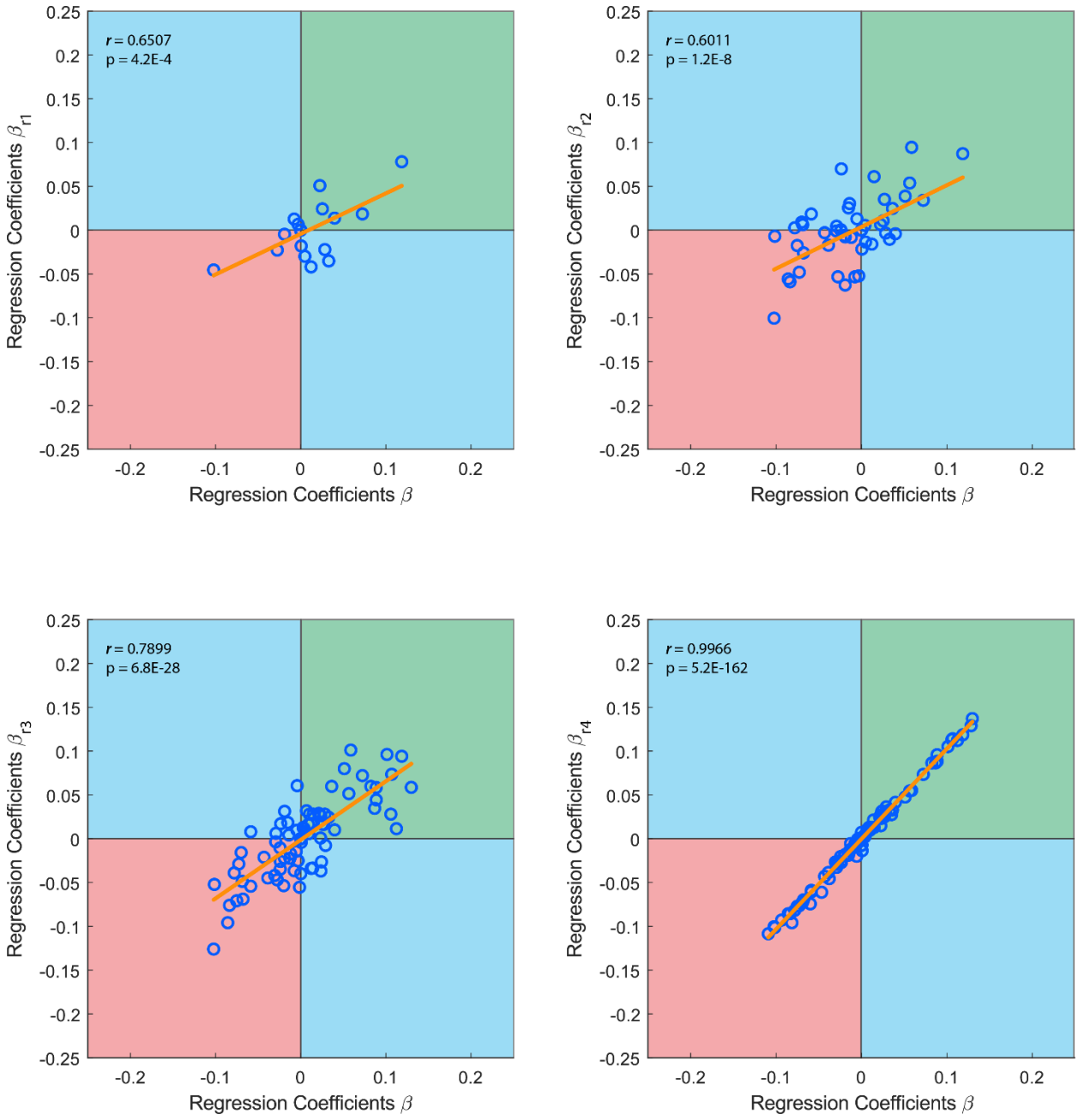



*Figure 61: Correlation between Regression Coefficients r = n and r < n for exponential growth rate fitness. The correlation (**r**) and its significance (p) between Regression Coefficients $\beta$ when r = n and r < n. When points are in the green, red, or blue squares, both coefficients indicate positive epistasis, negative epistasis, or have opposite signs respectively.*

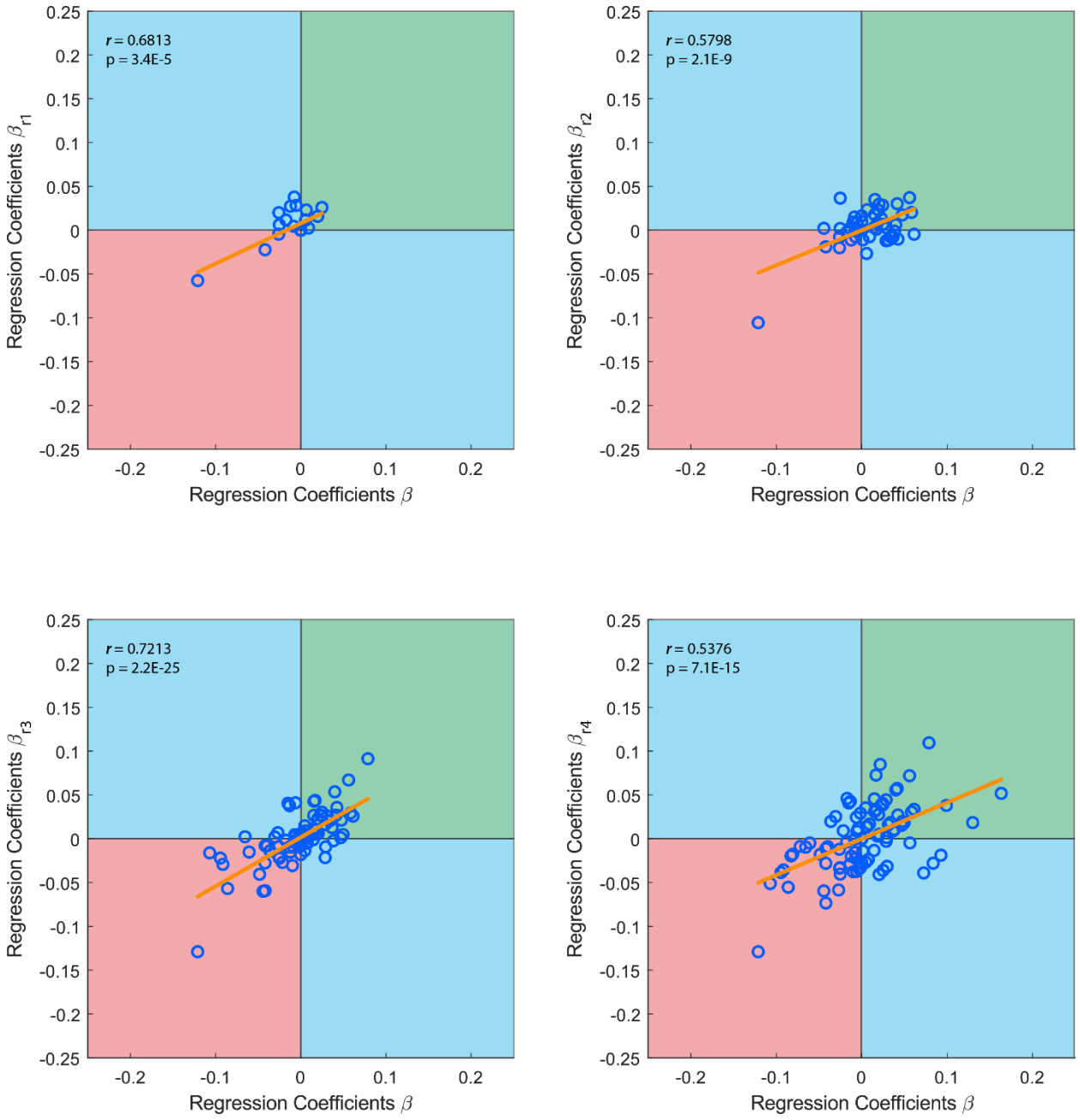



*Figure 62: Correlation between Regression Coefficients r = n and r < n for maximum OD after 20 hours fitness. The correlation (**r**) and its significance (p) between Regression Coefficients $\beta$ when r = n and r < n. When points are in the green, red, or blue squares, both coefficients indicate positive epistasis, negative epistasis, or have opposite signs respectively.*

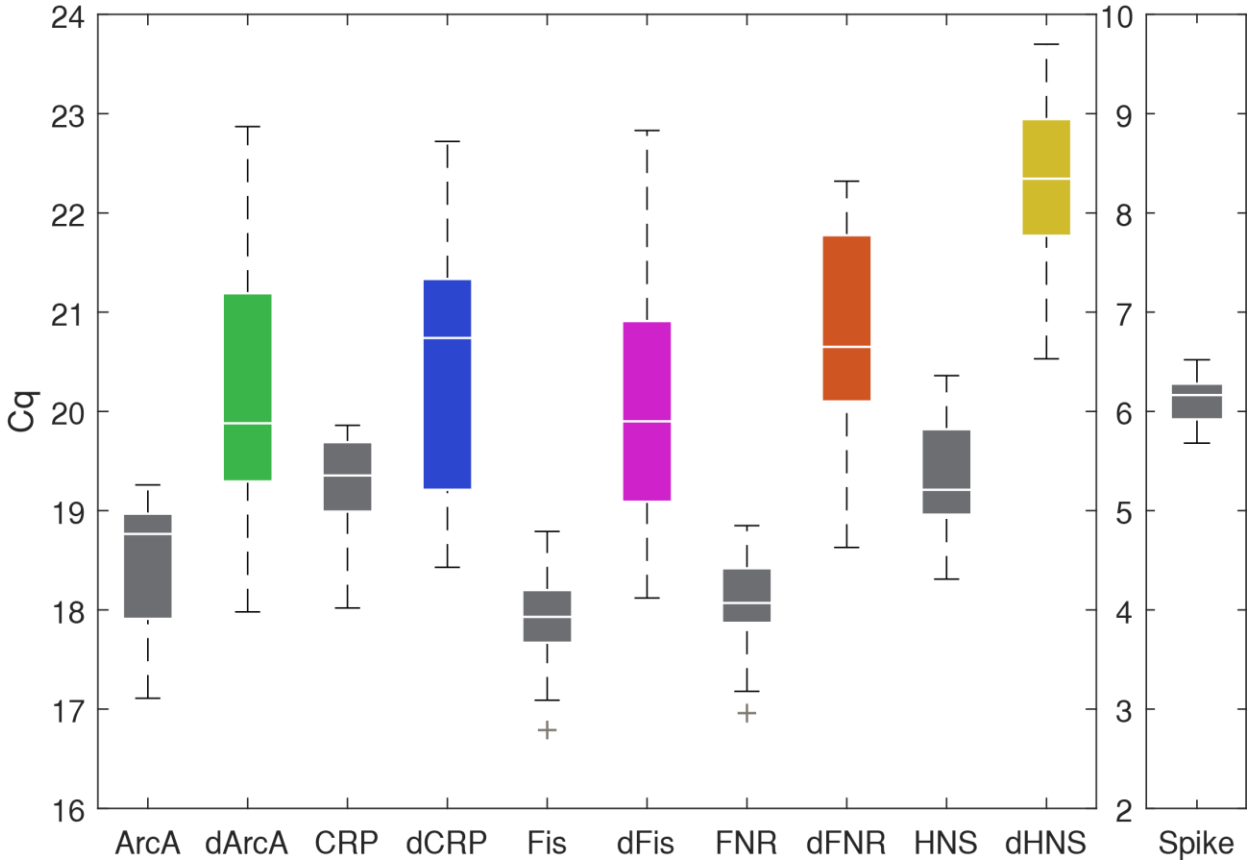



*Figure 63: Comparison of Genetic knock down strength in pCKDL strains using qRT-PCR. RNA was extracted from pCKDL strains at an OD of 0.3. An RNA spike was added to each RNA sample and qRT-PCR was performed for each gene targeted by the CRISPR system. The Cq was determined by regression and is down for strains containing either a targeting spacer for a given gene (coloured boxes; green: arcA, blue: crp, pink: fis, orange: fnr, yellow: hns) or a non-targeting spacer (grey boxes). Boxes represent the 25$^{th}$ and 75$^{th}$ percentile, the white line representing the median of the data, and whiskers represent the limits of the data not considered outliers. Outliers are plotted individually as grey '+'.*

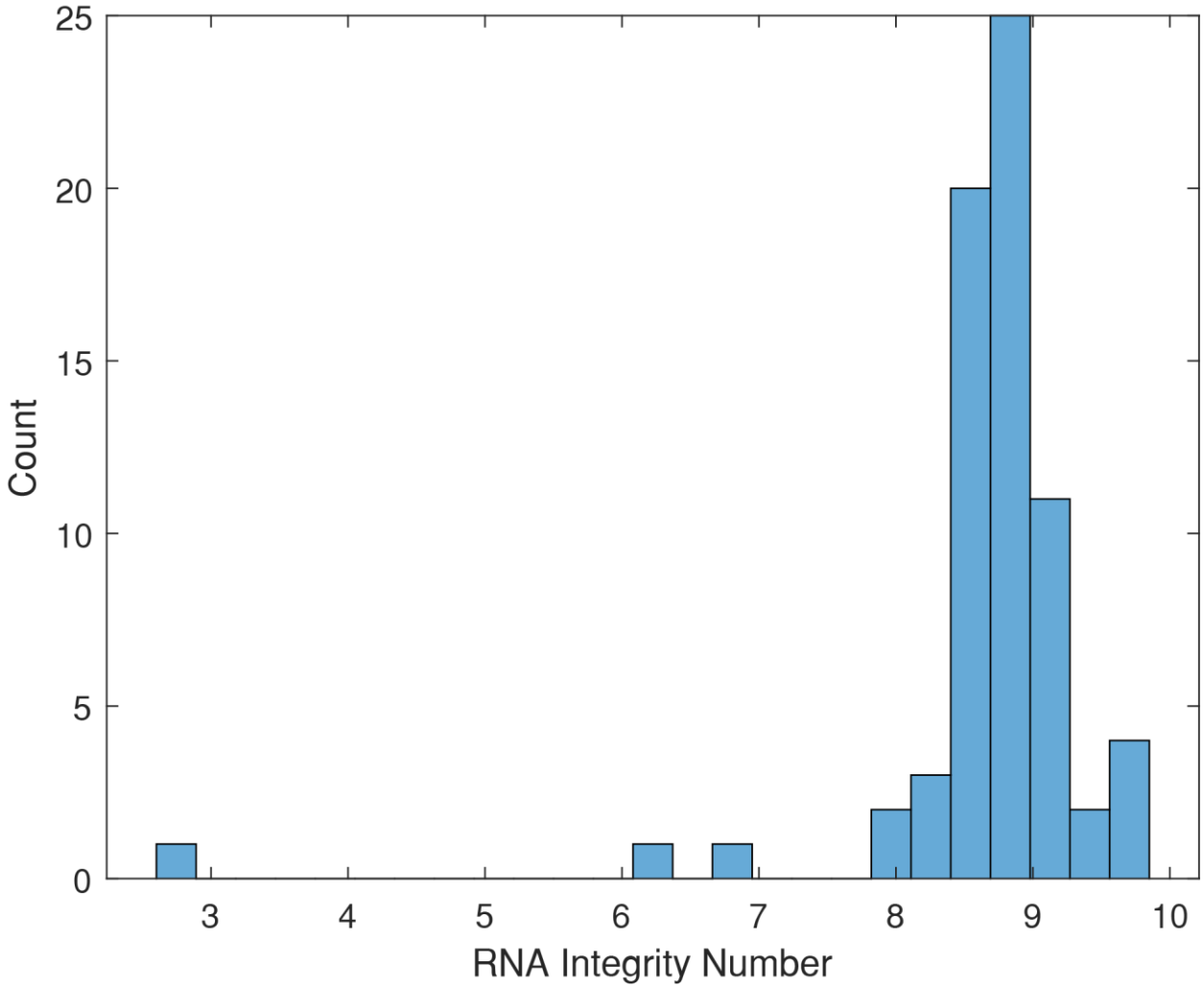



*Figure 64: RNA integrity of pCKDL samples for RNAtag-Seq. RNA was extracted for pCKDL strains grown in LB for 1.5 or 3.5 hours. RNA integrity (RIN) was recorded with Agilent TapeStation. Distribution of RIN is shown. Any samples with an RIN less than 7 were repeated.*

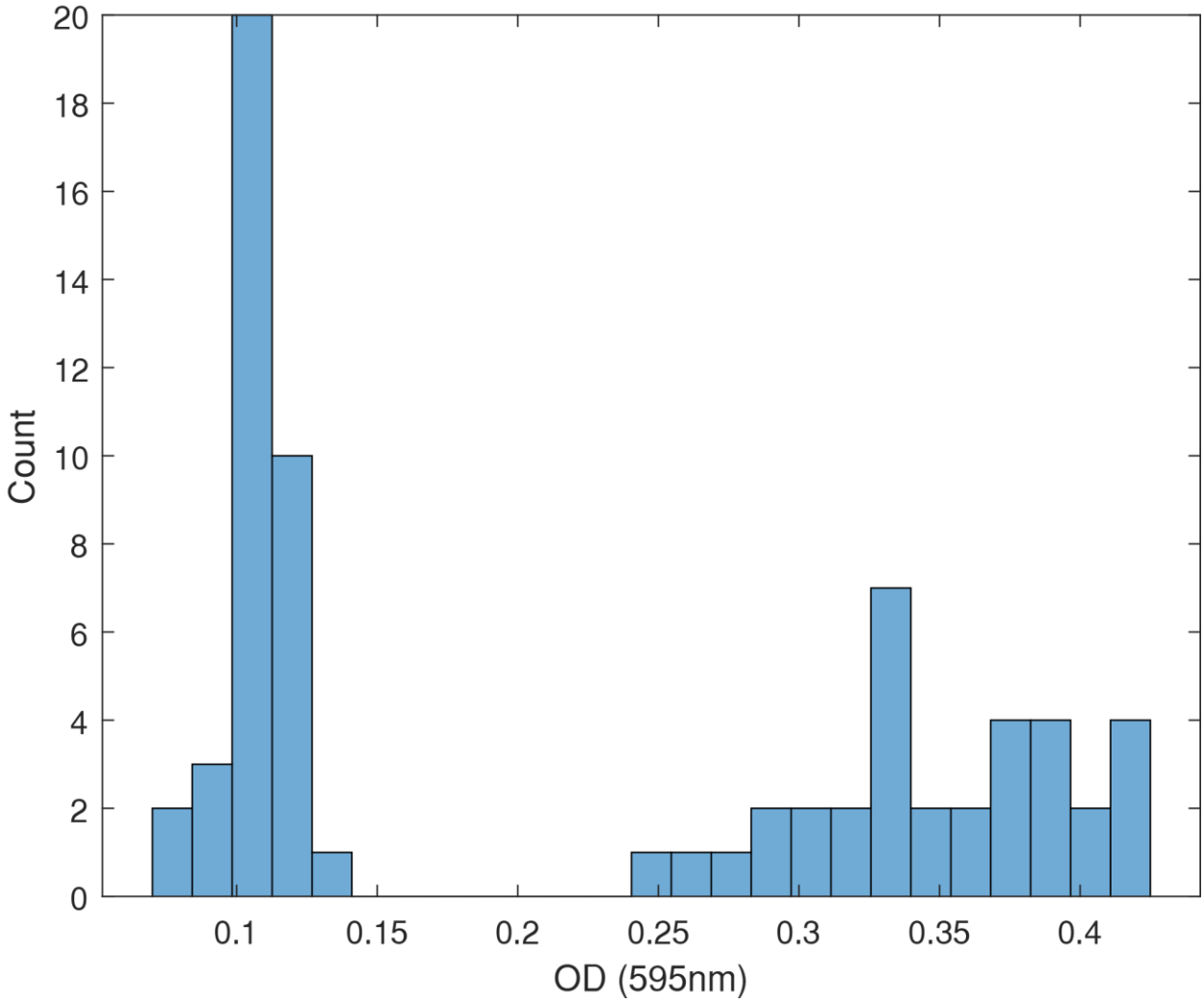



*Figure 65: Optical Density of pCKDL samples for RNAtag-Seq. Optical Density (OD) was measured at 595nm for pCKDL strains grown in LB for 1.5 or 3.5 hours. OD was measured on 200 µL of sample in 96 well plate. An OD of 0.1 ≈ 0.3 and 0.35 ≈ 1.3 when measure with a 1 cm cuvette. These correspond to mid exponential growth phase and early stationary phase respectively.*

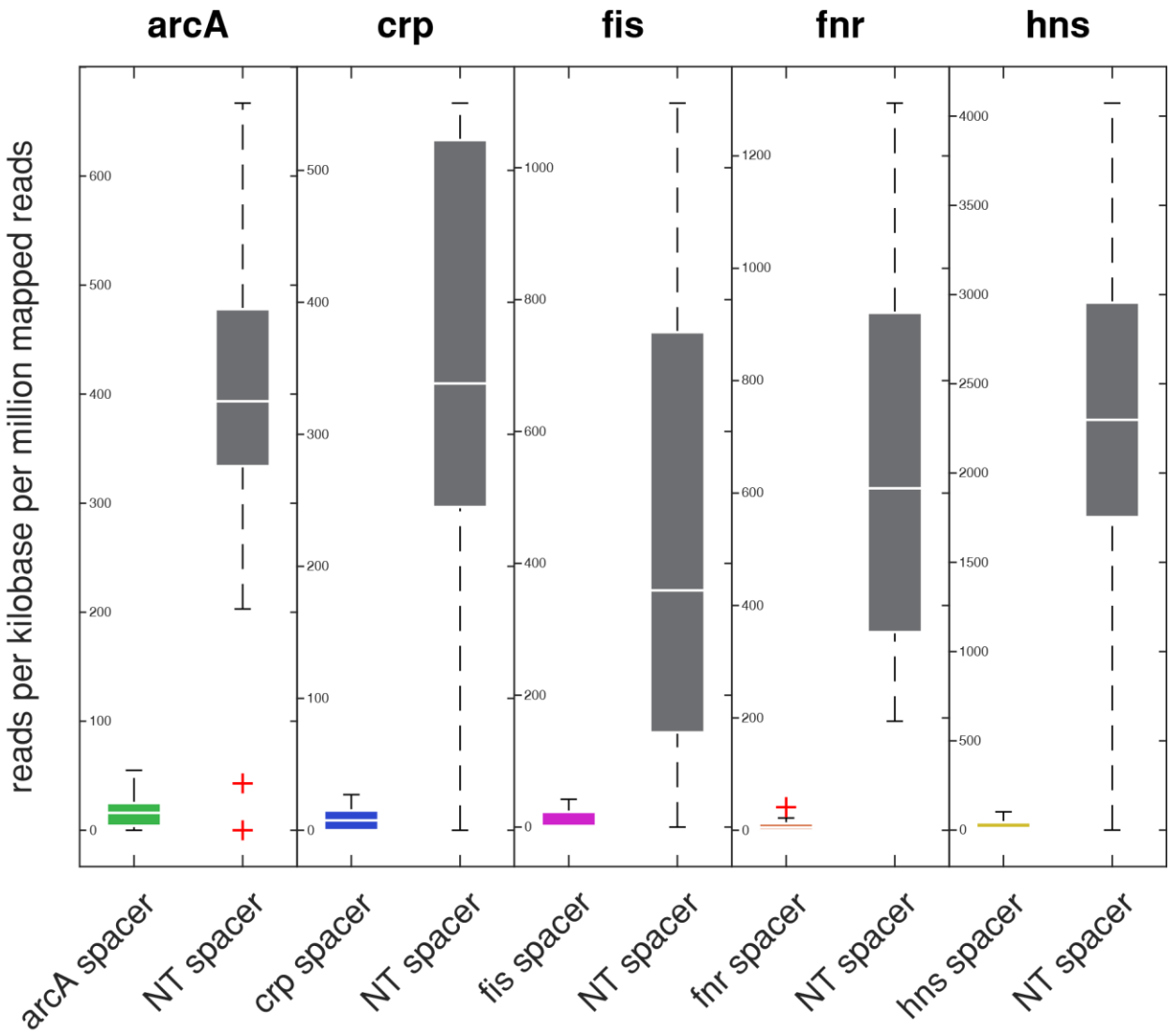



*Figure 66: CRISPR-Cas9 eliminates gene expression in perturbed strains. The quantity of RNA detected by RNAtag-seq for pCKDL strains that contain either a targeting spacer (coloured boxes; green: arcA, blue: crp, pink: fis, orange: fnr, yellow: hns) or a non-targeting spacer (grey boxes). Boxes represent the 25$^{th}$ and 75$^{th}$ percentile, the white line representing the median of the data, and whiskers represent the limits of the data not considered outliers. Outliers are plotted individually as red '+'.*

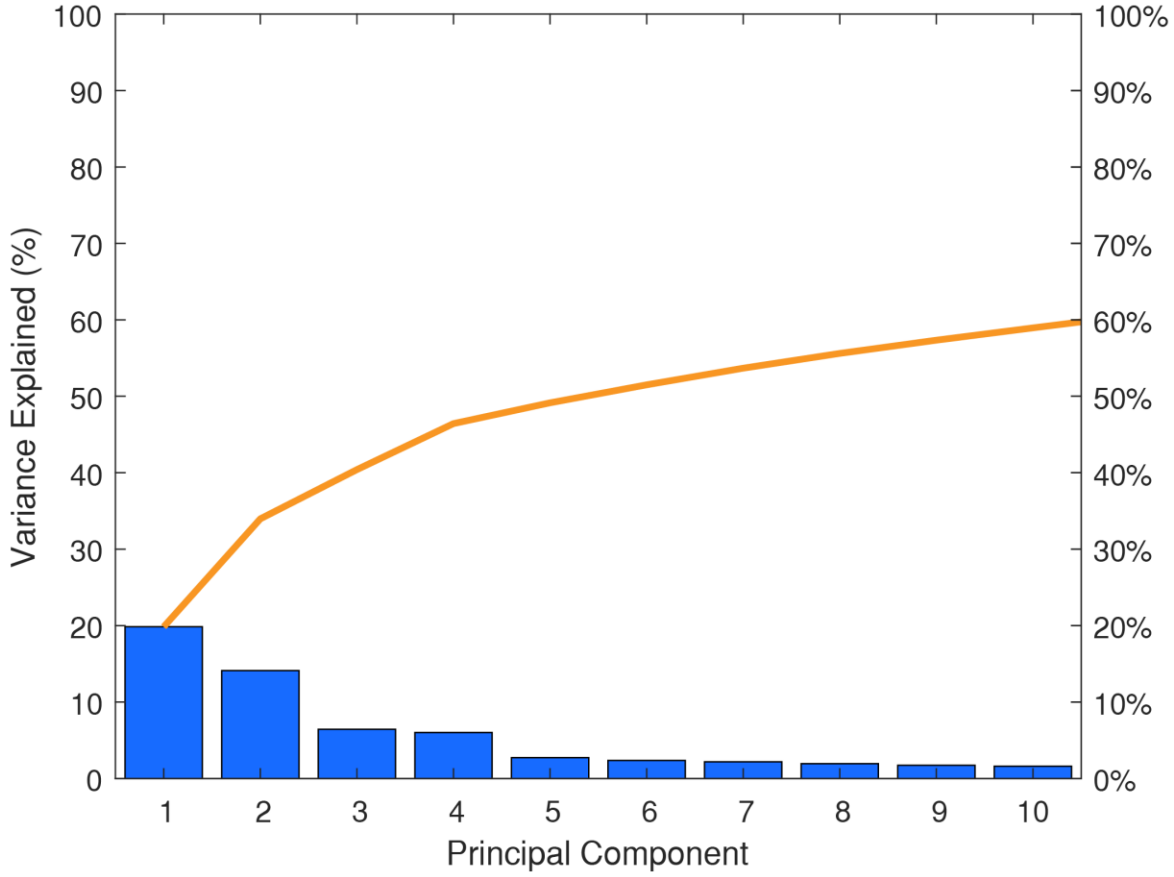



*Figure 67: Variance explained from each of the first 10 principle components in RNAtag-seq data. Each box represents the percentage of variance explained a given principle component in RNAtag-seq data from pCKDL strains grown in LB media. The orange line represents the cumulative variance explained by the principle components.*

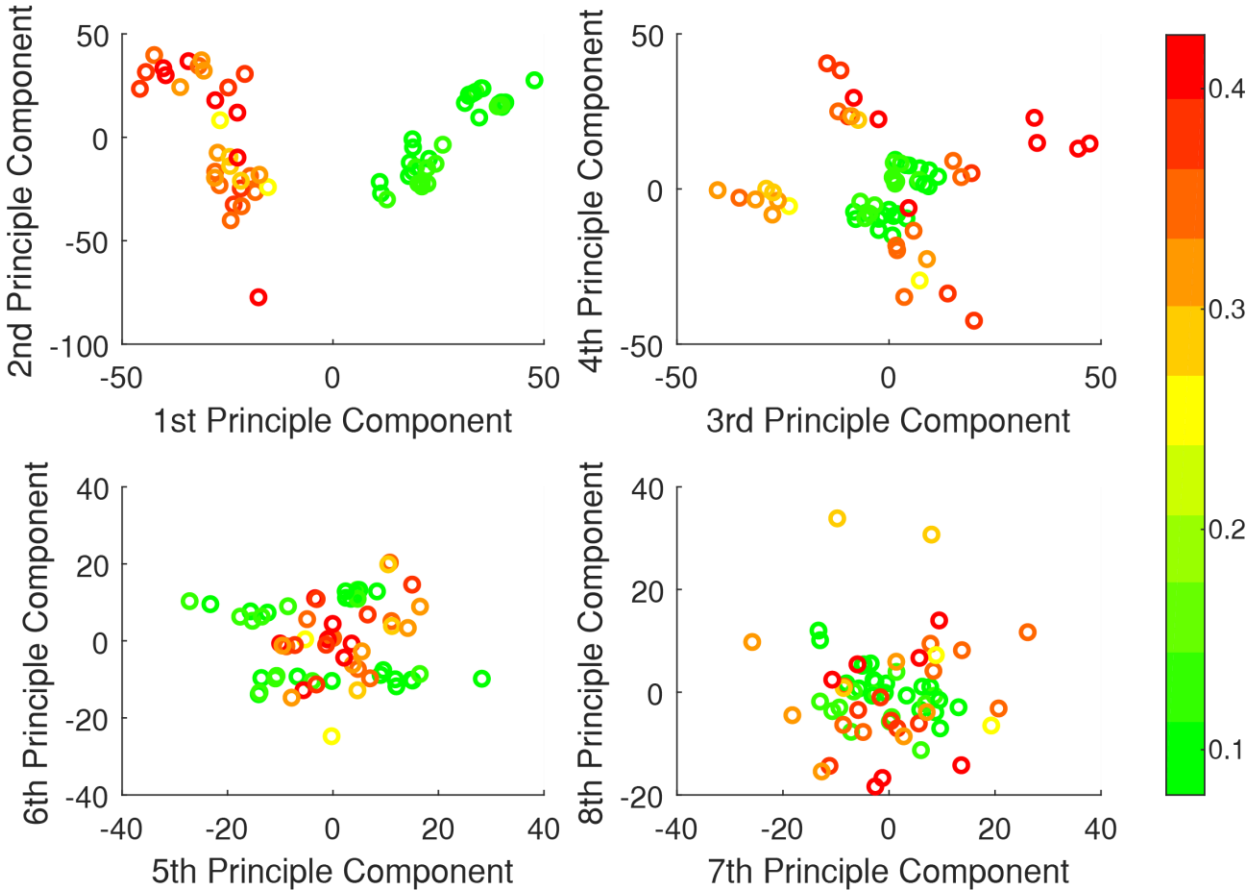



*Figure 68: Principle Components association to Optical Density. Transcription of pCKDL strains projected onto the first 8 principle components. Each circle represents the Principle component scores for a given pCKDL sample, coloured by the Optical Density of the strain when RNA was extracted.*

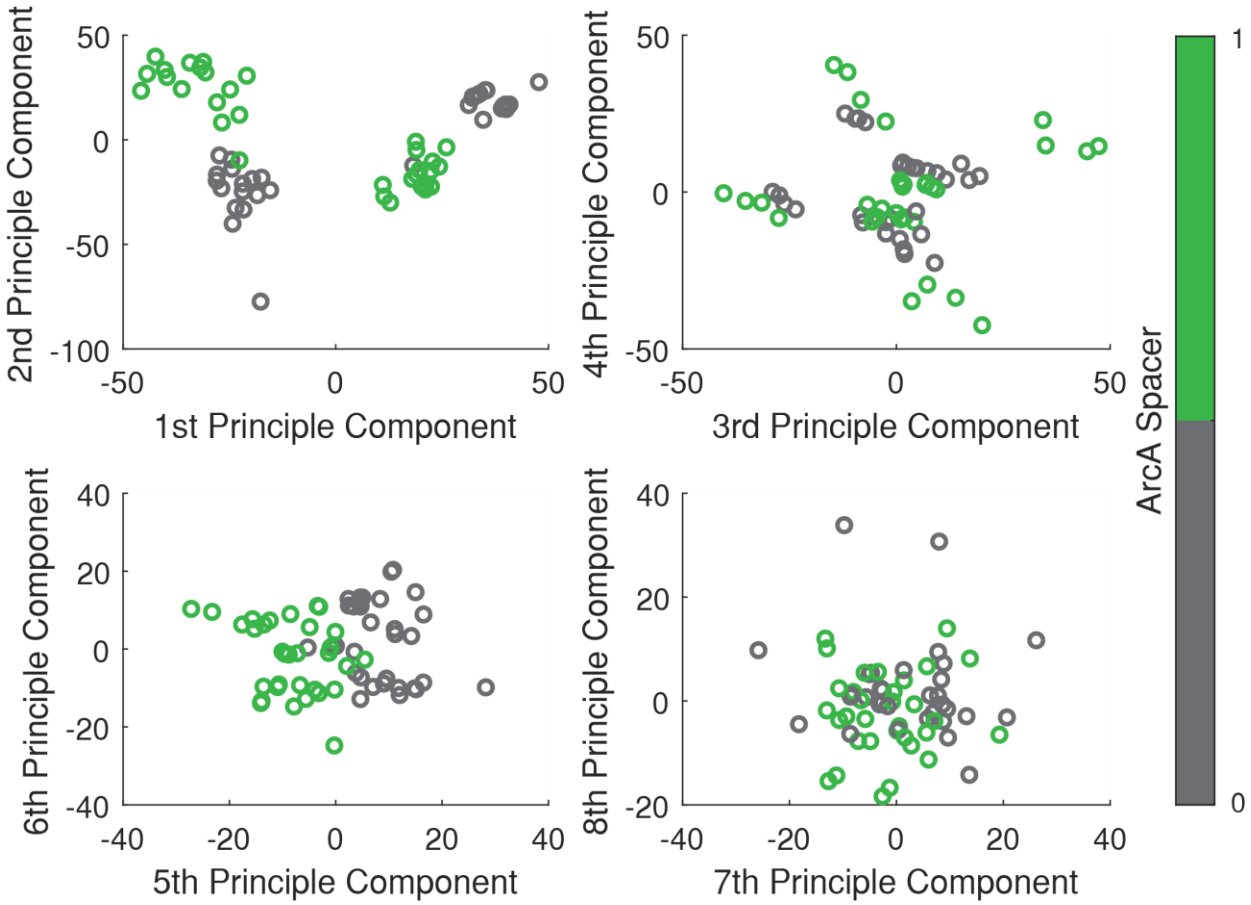



*Figure 69: Principle Components association to ArcA targeting spacer. Transcription of pCKDL strains projected onto the first 8 principle components. Each circle represents the Principle component scores for a given pCKDL sample, coloured by the presence (green) or absence (grey) of a ArcA targeting spacer.*

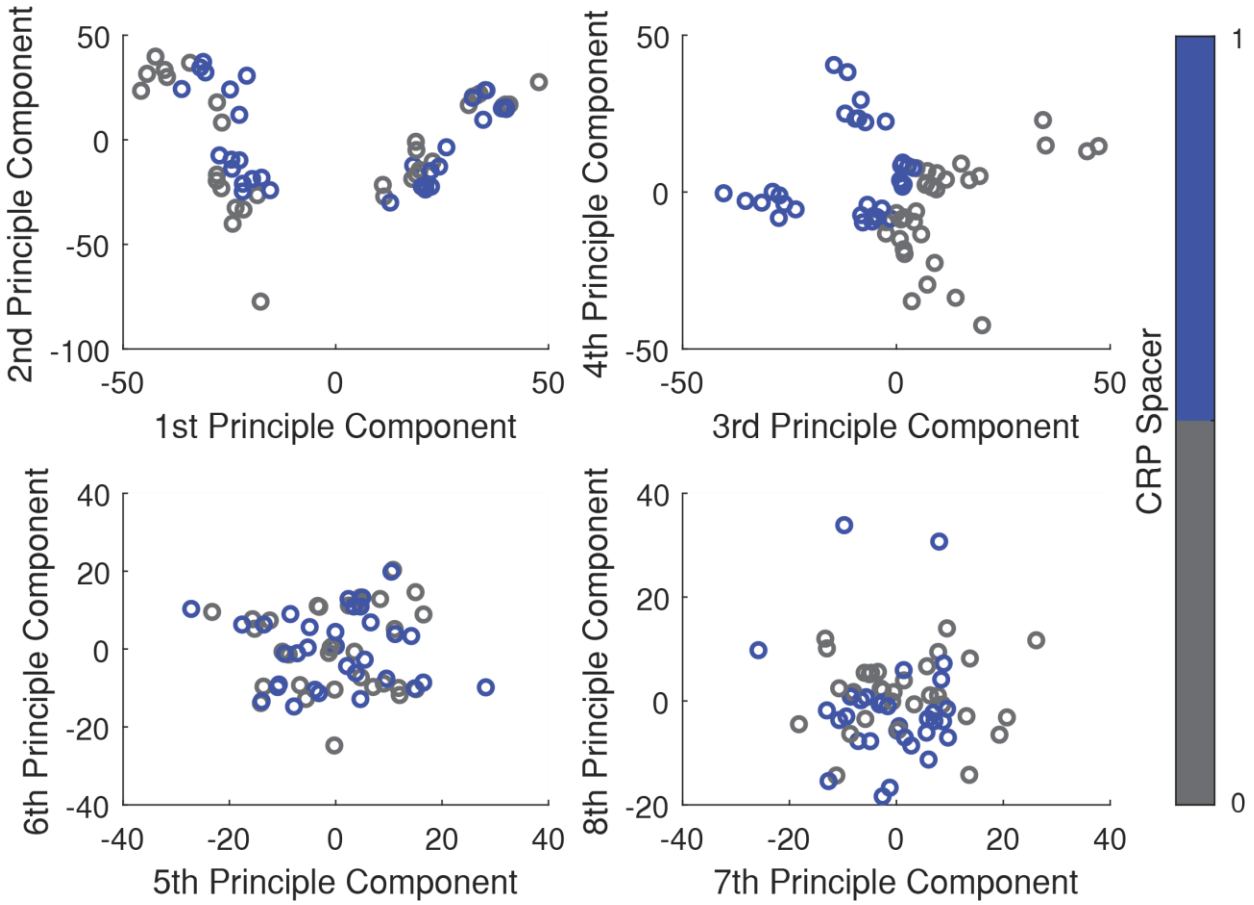



*Figure 70: Principle Components association to CRP targeting spacer. Transcription of pCKDL strains projected onto the first 8 principle components. Each circle represents the Principle component scores for a given pCKDL sample, coloured by the presence (blue) or absence (grey) of a CRP targeting spacer.*

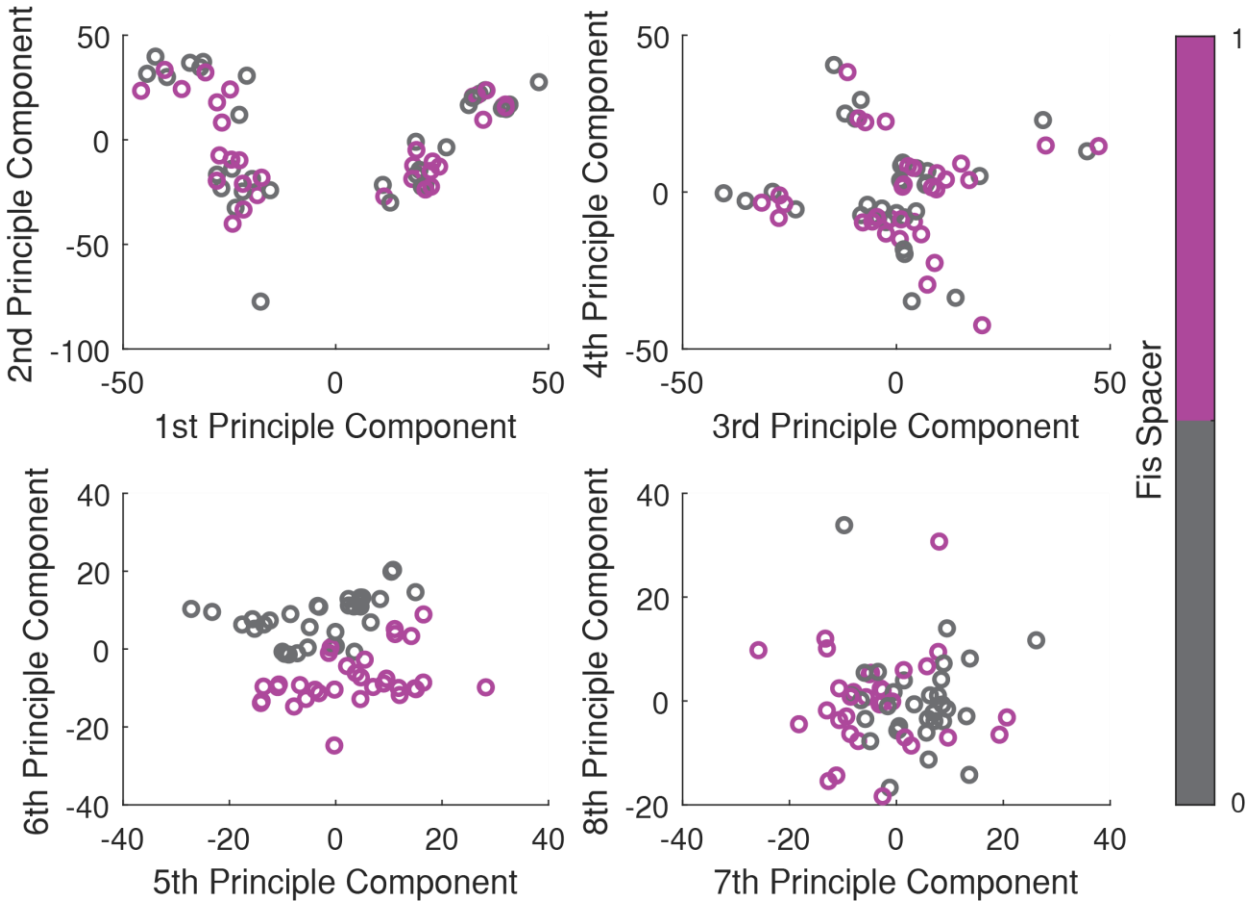



*Figure 71: Principle Components association to Fis targeting spacer. Transcription of pCKDL strains projected onto the first 8 principle components. Each circle represents the Principle component scores for a given pCKDL sample, coloured by the presence (pink) or absence (grey) of a Fis targeting spacer.*

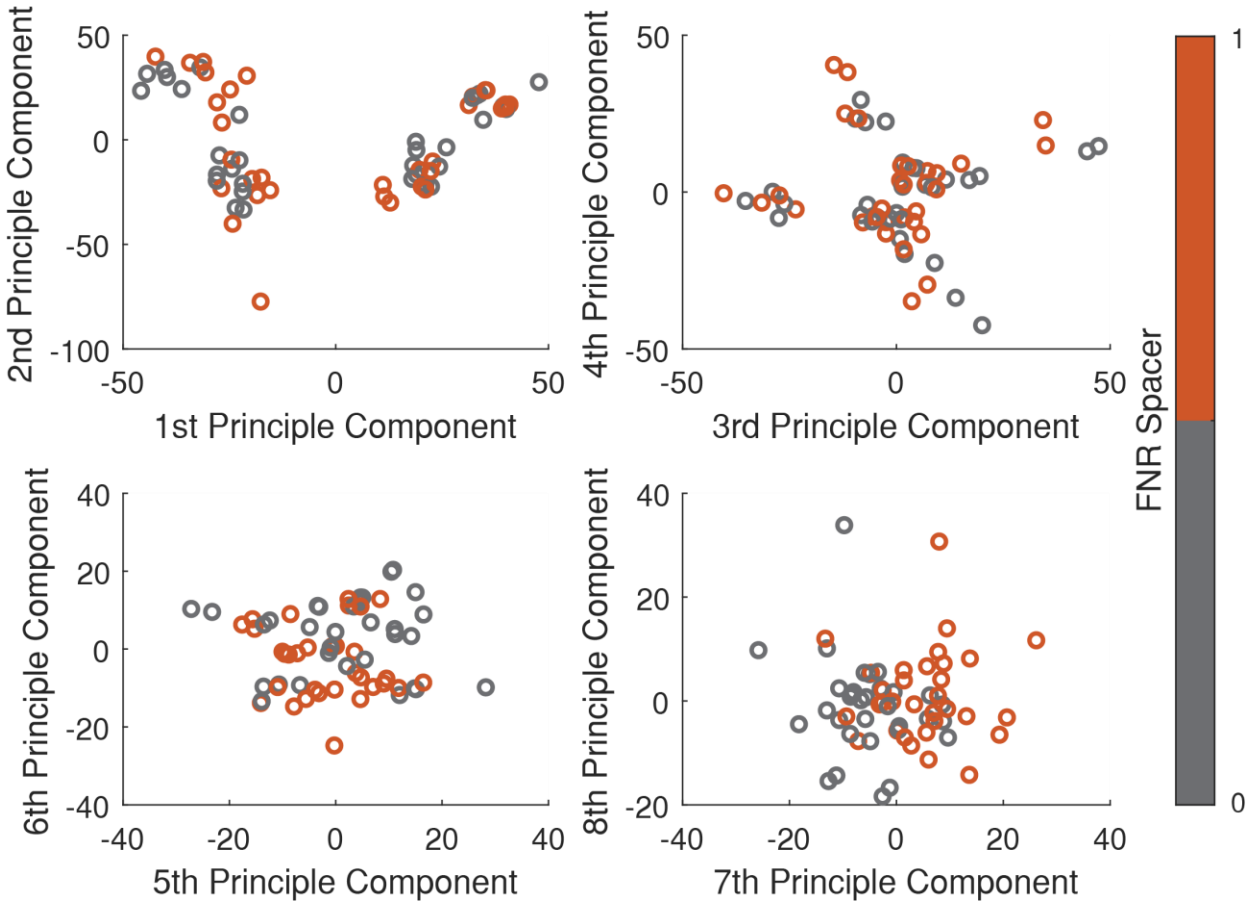



*Figure 72: Principle Components association to FNR targeting spacer. Transcription of pCKDL strains projected onto the first 8 principle components. Each circle represents the Principle component scores for a given pCKDL sample, coloured by the presence (orange) or absence (grey) of a FNR targeting spacer.*

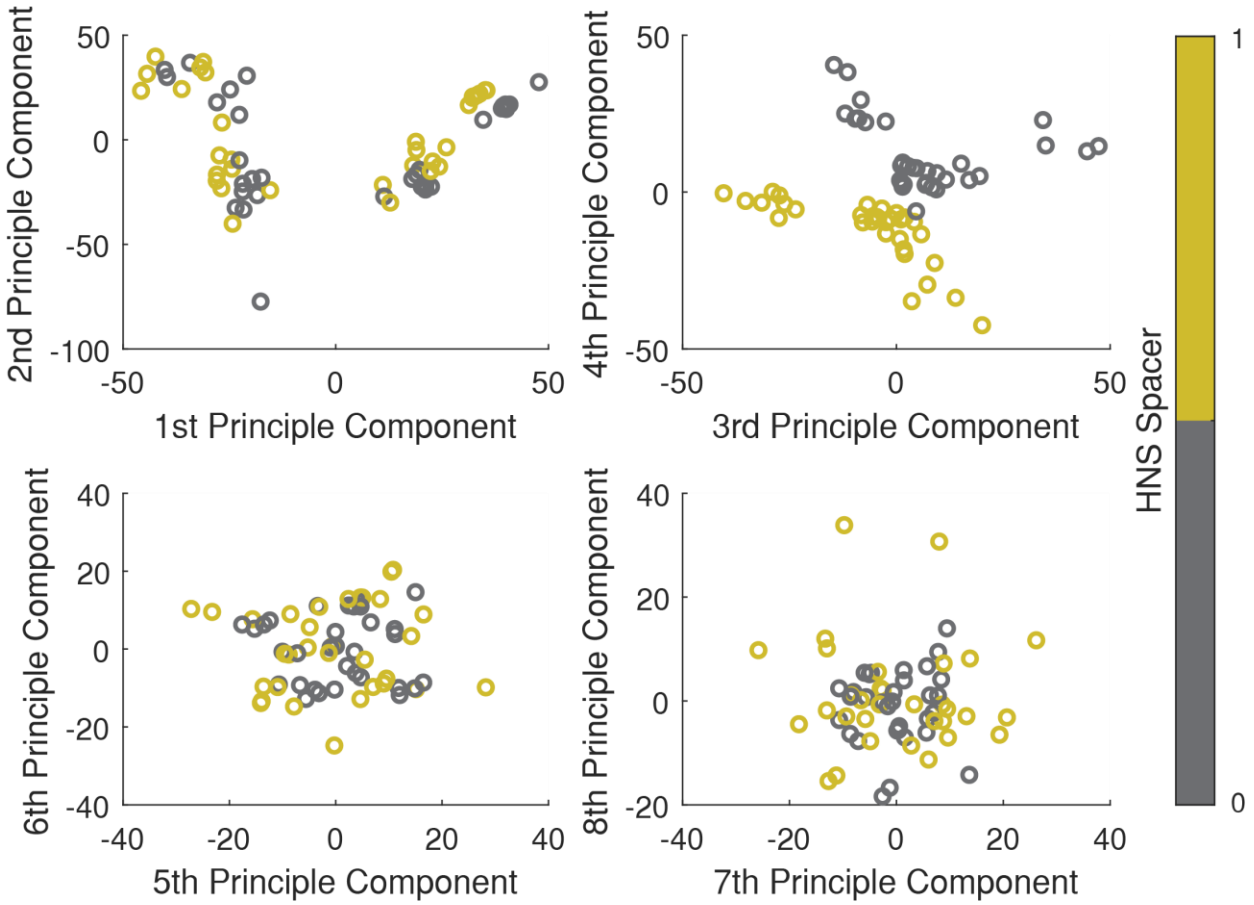



*Figure 73: Principle Components association to HNS targeting spacer. Transcription of pCKDL strains projected onto the first 8 principle components. Each circle represents the Principle component scores for a given pCKDL sample, coloured by the presence (yellow) or absence (grey) of a HNS targeting spacer.*

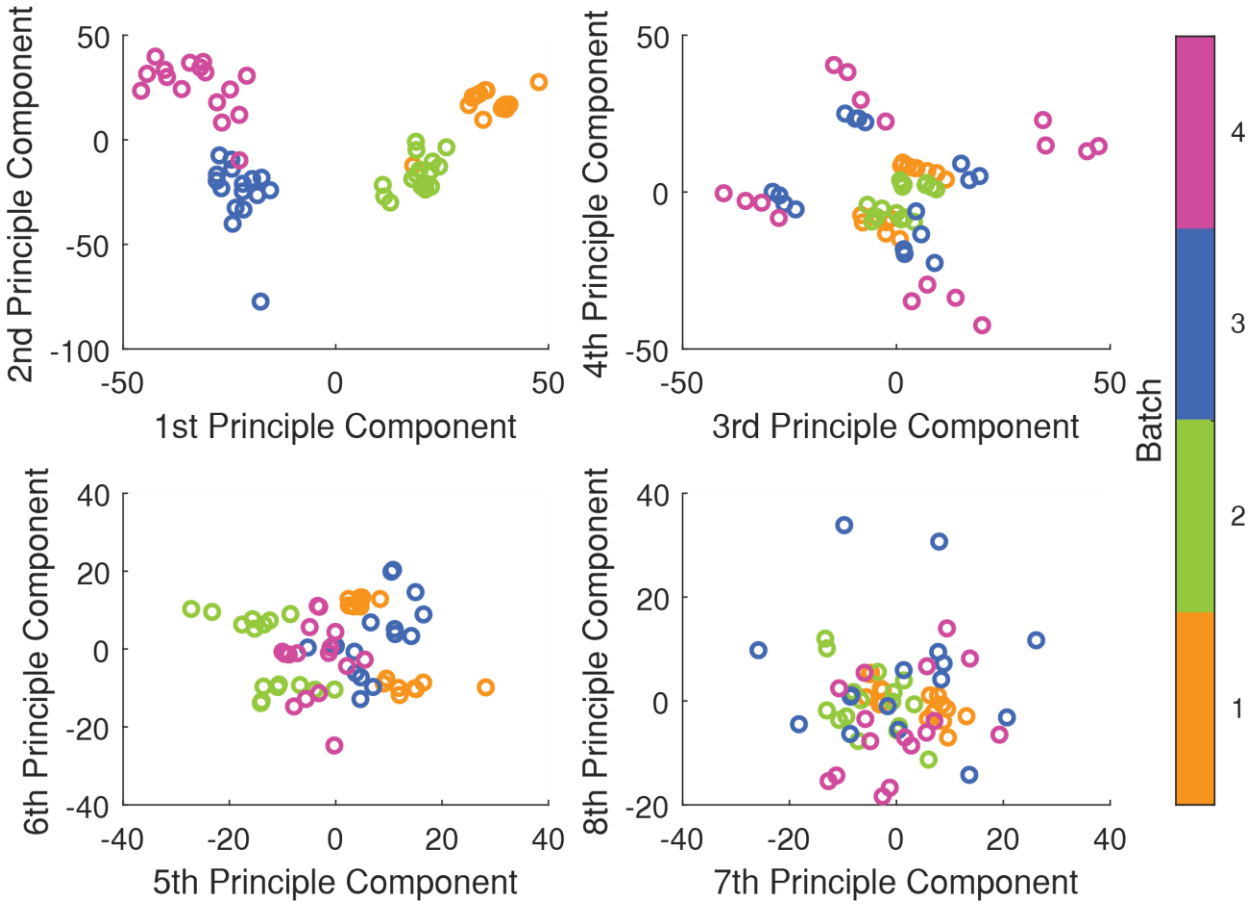



*Figure 74 Principle Components association to RNAtag-seq Batch. Transcription of pCKDL strains projected onto the first 8 principle components. Each circle represents the Principle component scores for a given pCKDL sample, coloured by the batch (pooled samples of RNA) they were contained in from RNAtag-seq protocol. Indicates any experimental biases.*

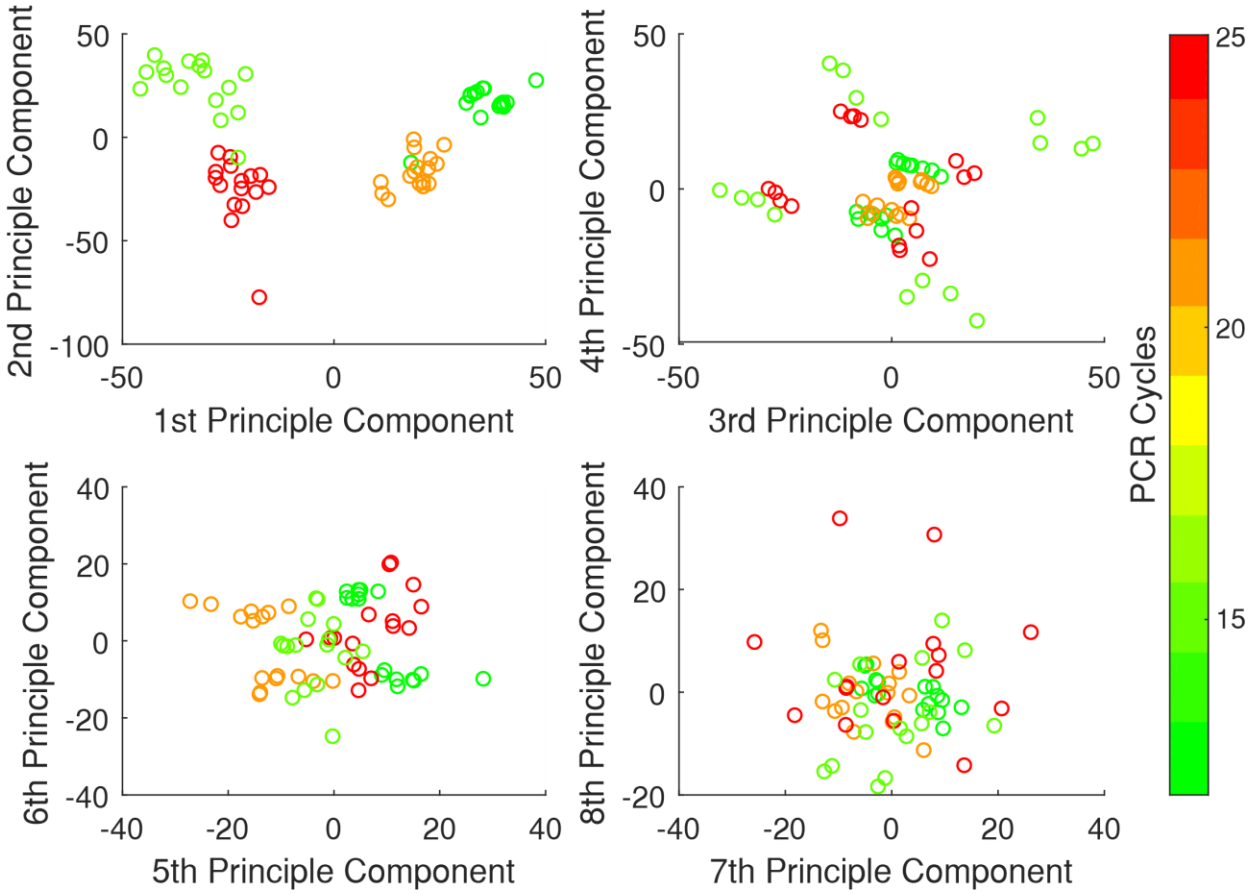



*Figure 75: Principle Components association to number of PCR cycles required before sequencing. Transcription of pCKDL strains projected onto the first 8 principle components. Each circle represents the Principle component scores for a given pCKDL sample, coloured by number of PCR cycles the sample underwent before sequencing. Indicates biases from RT efficiency, single strand DNA ligation efficiency, and PCR amplification biases.*

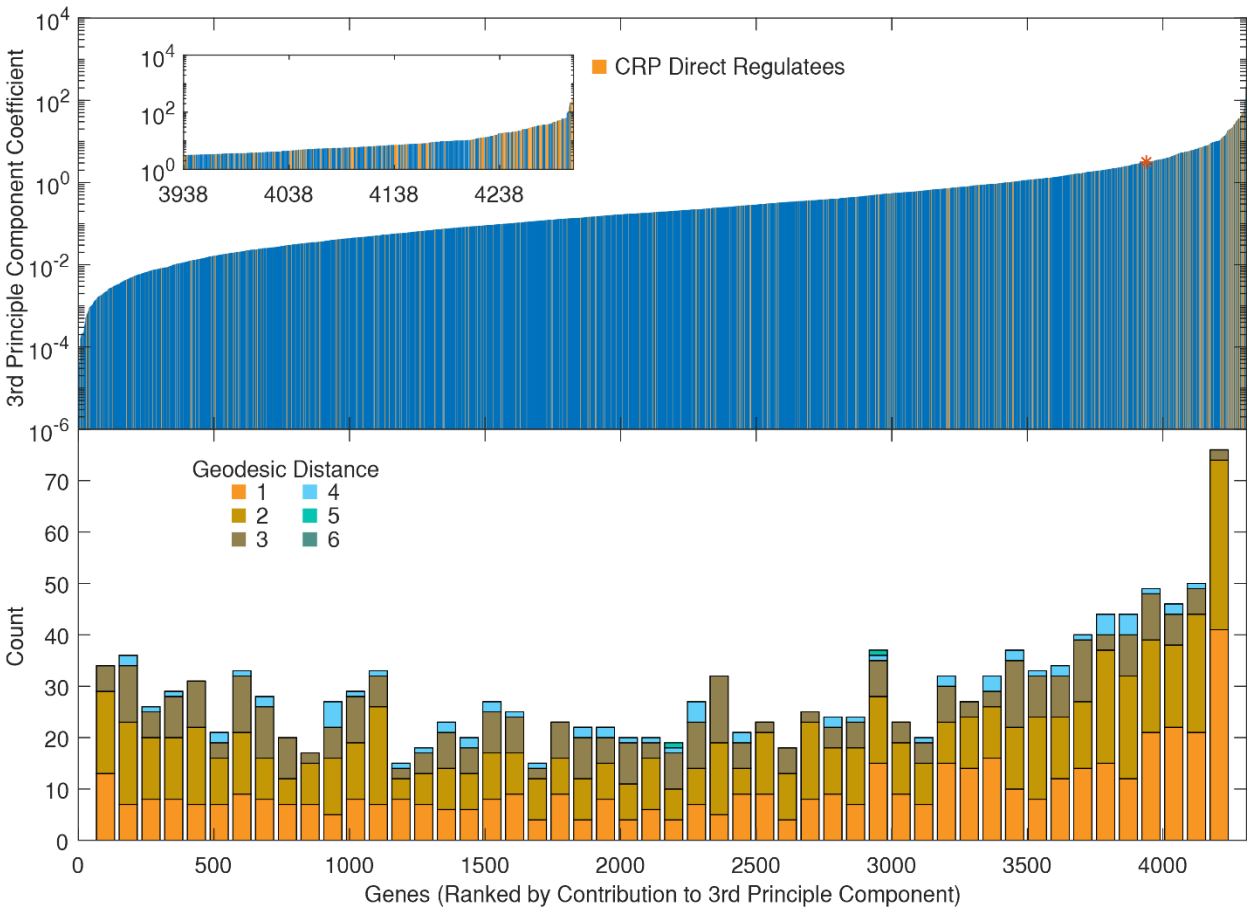



*Figure 76: Mapping of direct and indirect regulatees of CRP onto genes sorted by their contribution to the 3rd Principle Component. Top: Genes' contribution to the 3rd principle component. Genes highlighted in yellow are directly regulated by CRP. The star on the curve represents contributions above 1% of the maximum. Insert is expanded view of all the genes above this threshold. Bottom: Genes are separated into 50 equal sized bins. The number of genes in each bin which are directly regulated by CRP (Geodesic Distance of 1) are shown in yellow. The number of genes in each bin which are indirectly regulated by CRP (Geodesic Distance 2 to 6) are also shown.*

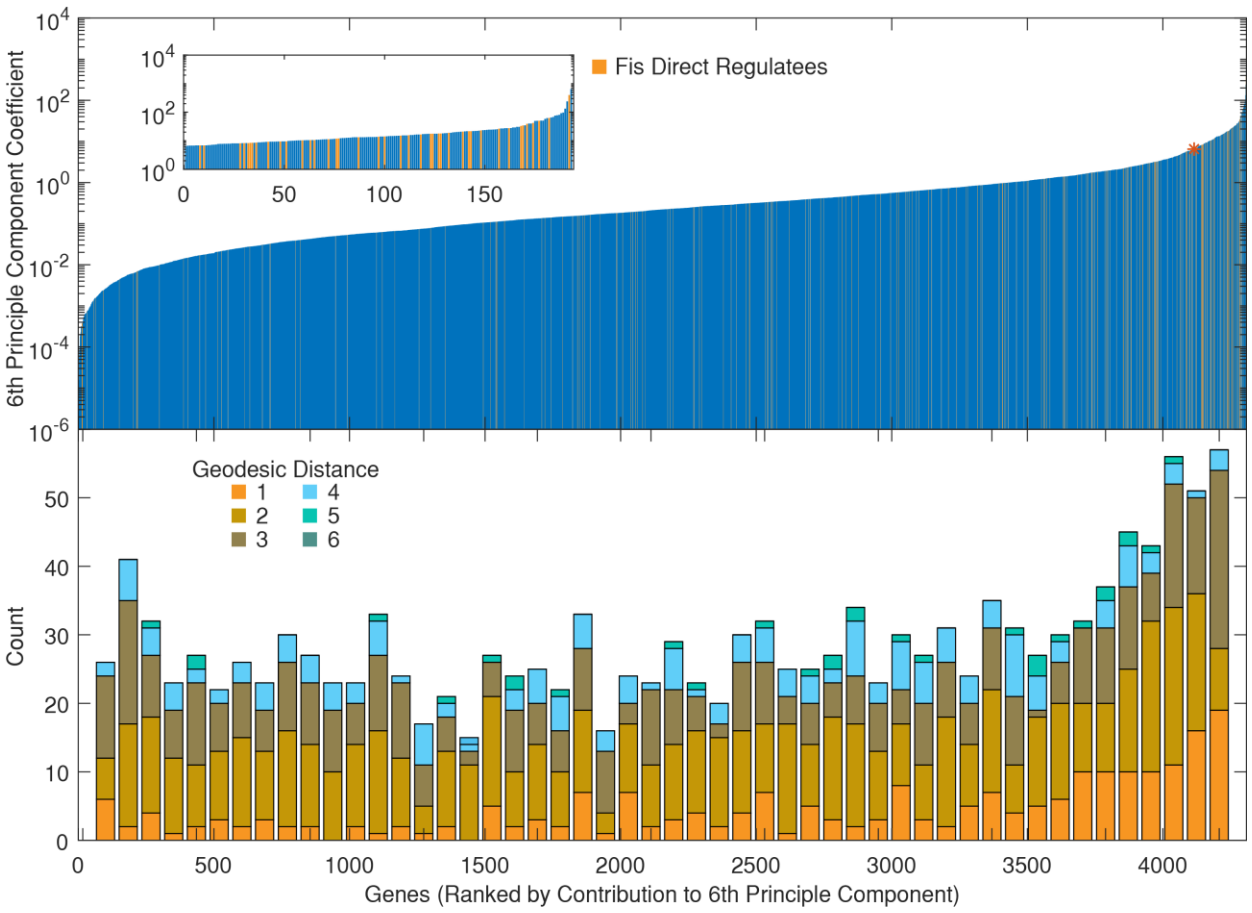



*Figure 77: Mapping of direct and indirect regulatees of Fis onto genes sorted by their contribution to the 6th Principle Component. Top: Genes' contribution to the 6th principle component. Genes highlighted in yellow are directly regulated by Fis. The star on the curve represents contributions above 1% of the maximum. Insert is expanded view of all the genes above this threshold. Bottom: Genes are separated into 50 equal sized bins. The number of genes in each bin which are directly regulated by Fis (Geodesic Distance of 1) are shown in yellow. The number of genes in each bin which are indirectly regulated by Fis (Geodesic Distance 2 to 6) are also shown.*

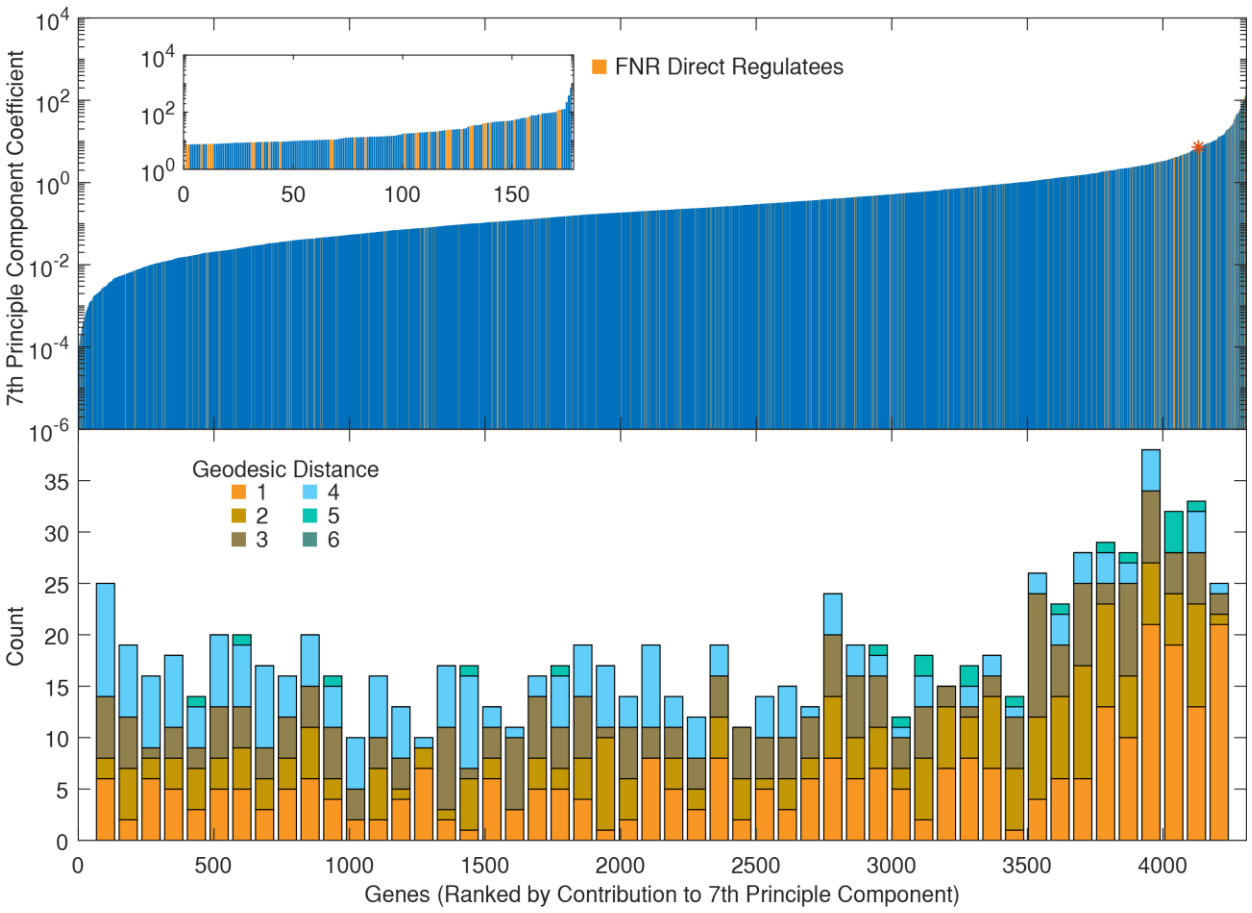



*Figure 78: Mapping of direct and indirect regulatees of FNR onto genes sorted by their contribution to the 7th Principle Component. Top: Genes' contribution to the 7th principle component. Genes highlighted in yellow are directly regulated by FNR. The star on the curve represents contributions above 1% of the maximum. Insert is expanded view of all the genes above this threshold. Bottom: Genes are separated into 50 equal sized bins. The number of genes in each bin which are directly regulated by FNR (Geodesic Distance of 1) are shown in yellow. The number of genes in each bin which are indirectly regulated by FNR (Geodesic Distance 2 to 6) are also shown.*

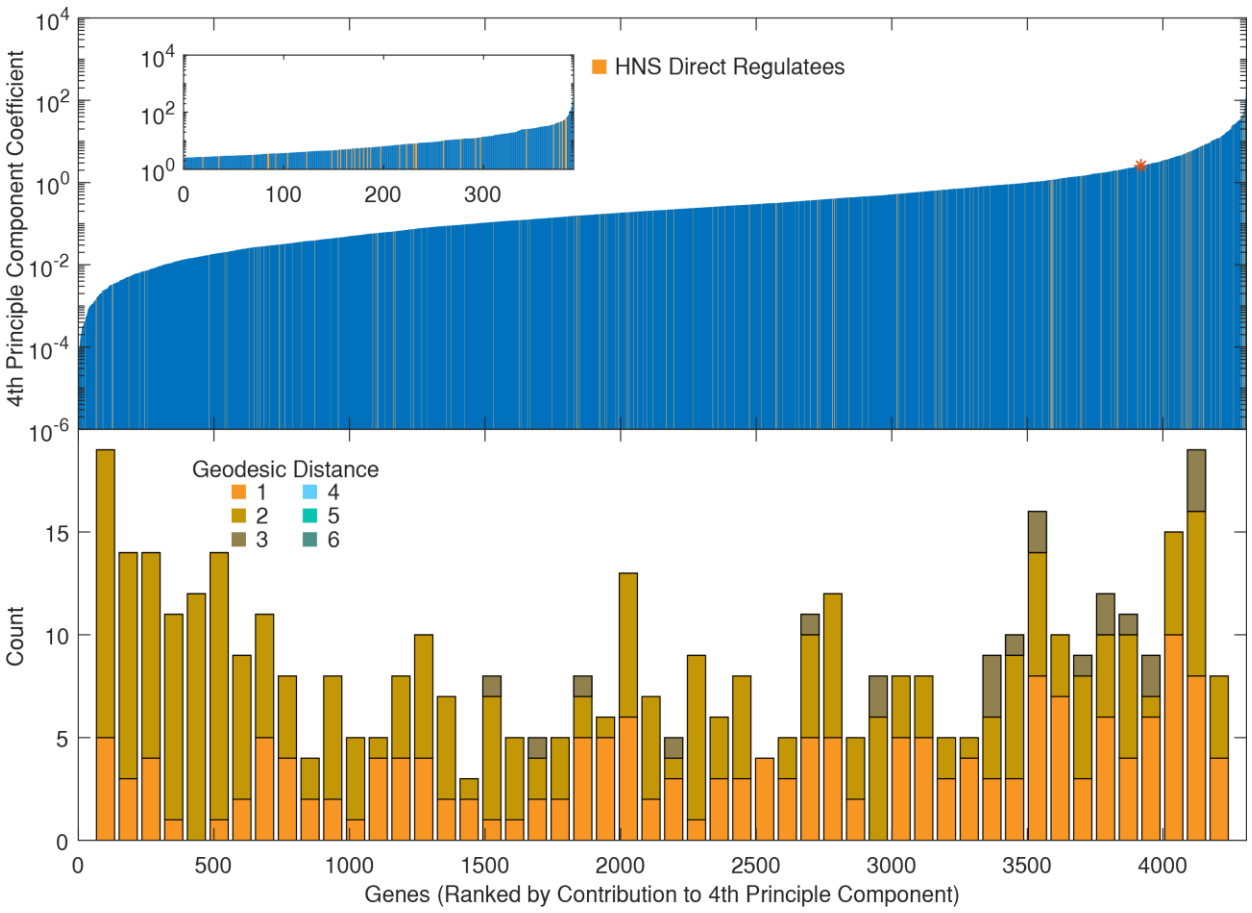



*Figure 79: Mapping of direct and indirect regulatees of HNS onto genes sorted by their contribution to the 4th Principle Component. Top: Genes' contribution to the 4th principle component. Genes highlighted in yellow are directly regulated by HNS. The star on the curve represents contributions above 1% of the maximum. Insert is expanded view of all the genes above this threshold. Bottom: Genes are separated into 50 equal sized bins. The number of genes in each bin which are directly regulated by HNS (Geodesic Distance of 1) are shown in yellow. The number of genes in each bin which are indirectly regulated by HNS (Geodesic Distance 2 to 6) are also shown.*

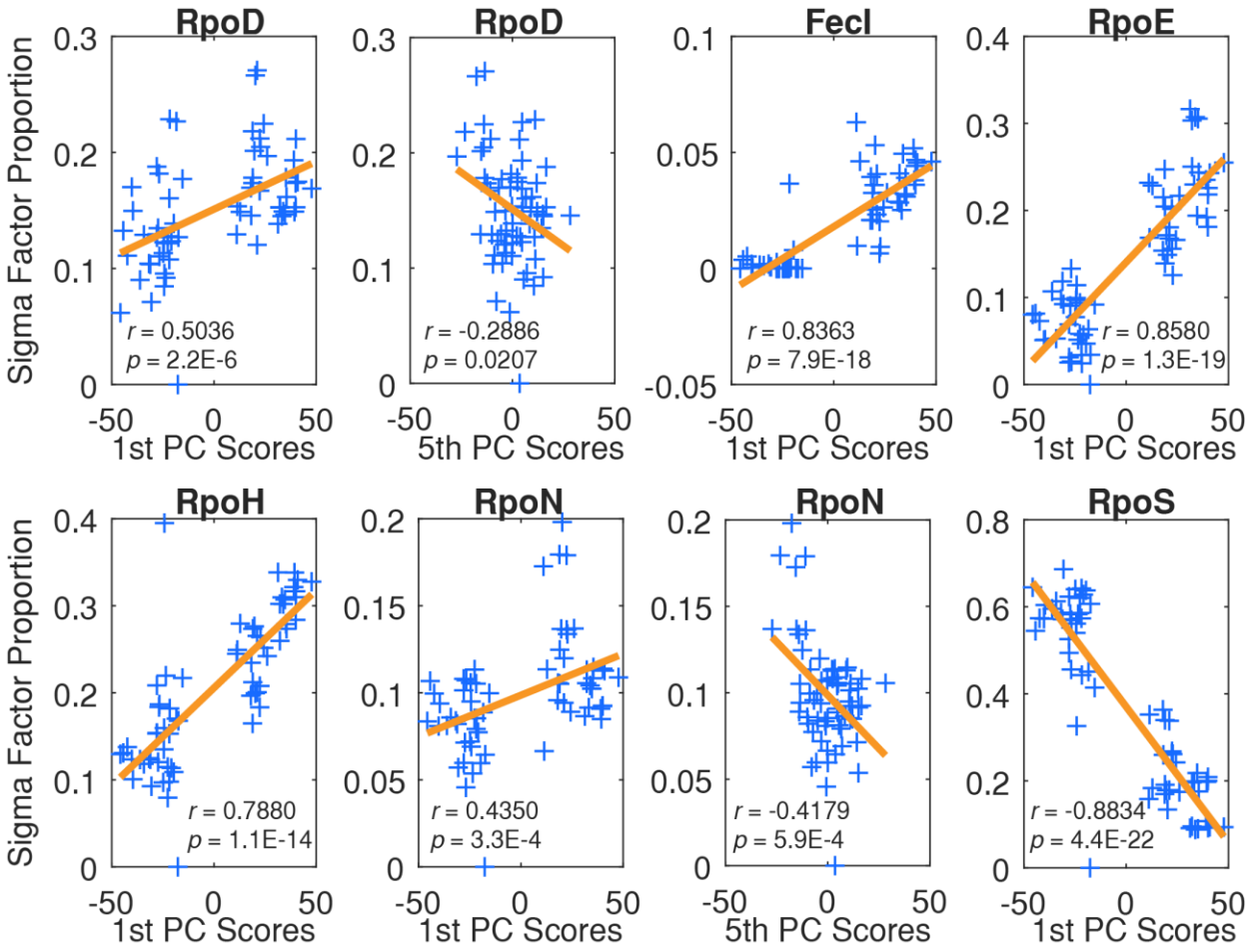



*Figure 80: Association of Sigma Factors to Principle Components. The proportion RNA detected from a given sigma factors from the total RNA detected for all sigma factors is plotted against the principle component scores for significantly ($p < 0.05$) correlated principle components. The correlation (r) and the significance (p) are shown for each plot.*

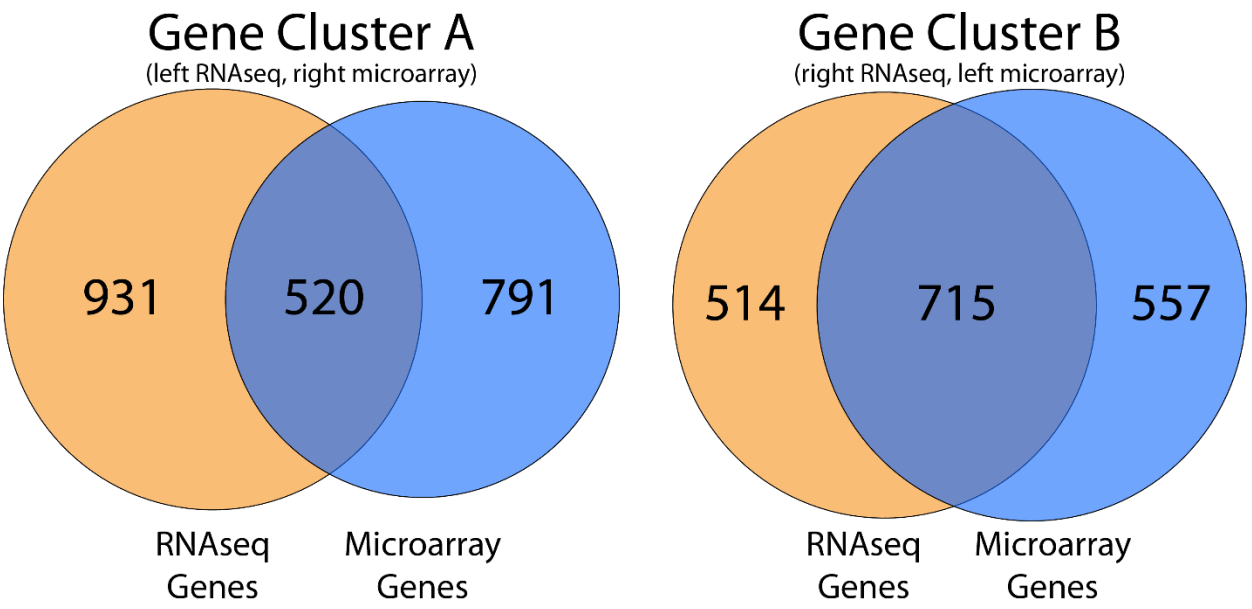



*Figure 81: Overlap of genes in each cluster from the first principle component between pCKDL and Many Microbe Microarrays Database. The number of genes unique to either RNAseq from pCKDL strains, or from Microarray data in the Many Microbe Microarray Database (M3D), are shown for each cluster of genes found in the first principle component. Since pCKDL data is anti-correlated with optical density and M3D is positively correlated, the clusters from opposite sides (right to left) of the first principle component are compared.*

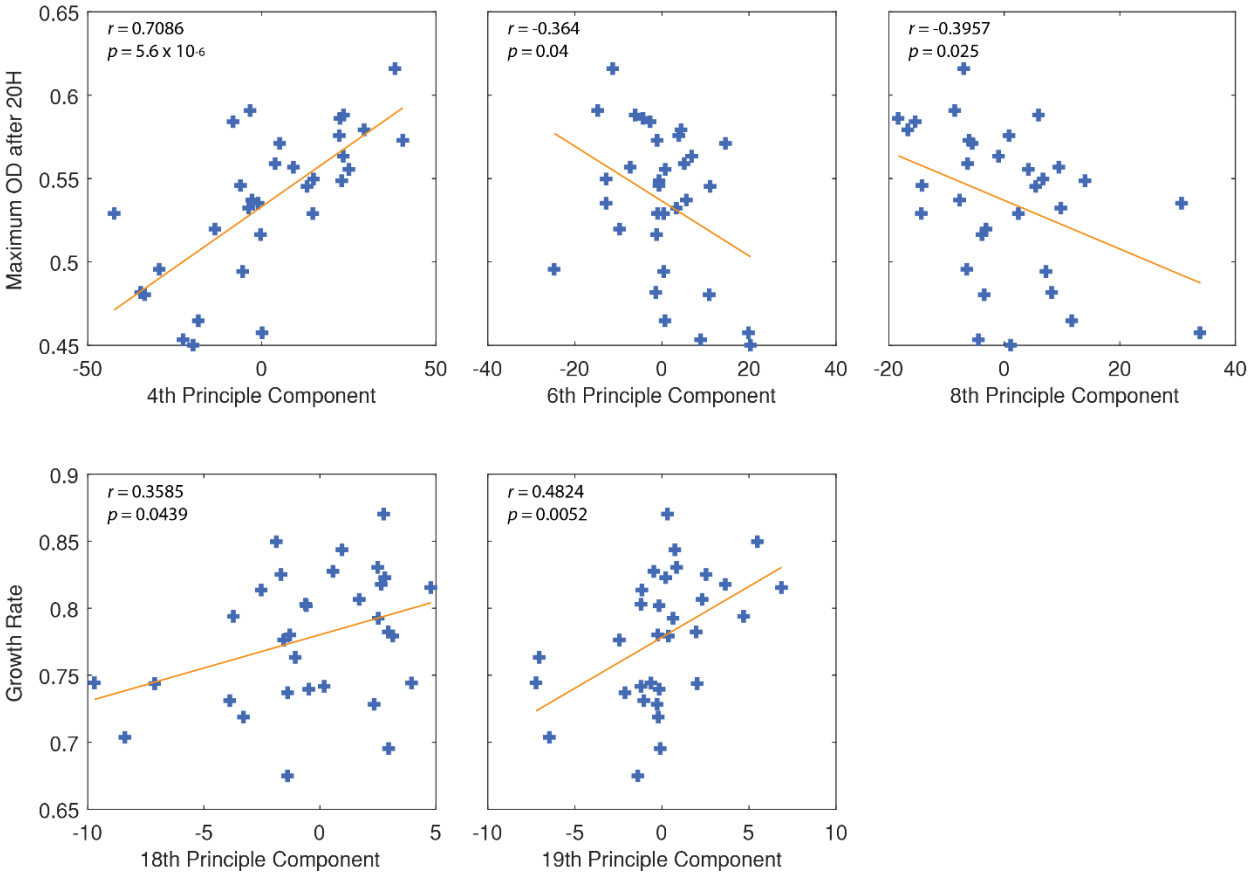



*Figure 82: Association of Fitness scores with principle components. The Growth Rate Fitness is plotted against the Principle Component scores for the pCKDL samples taken in exponential phase. The Maximum OD fitness is plotted against the Principle Component scores for the pCKDL samples taken in early stationary phase. Maximum OD fitness is strongly correlated with the 4$^{th}$ principle component (associated with HNS) and weakly anti-correlated with the 6$^{th}$ principle component (associated with Fis) and the 8$^{th}$ Principle component. Growth Rate fitness does not have any significant correlations to any principle components until the 18$^{th}$ and 19$^{th}$.*

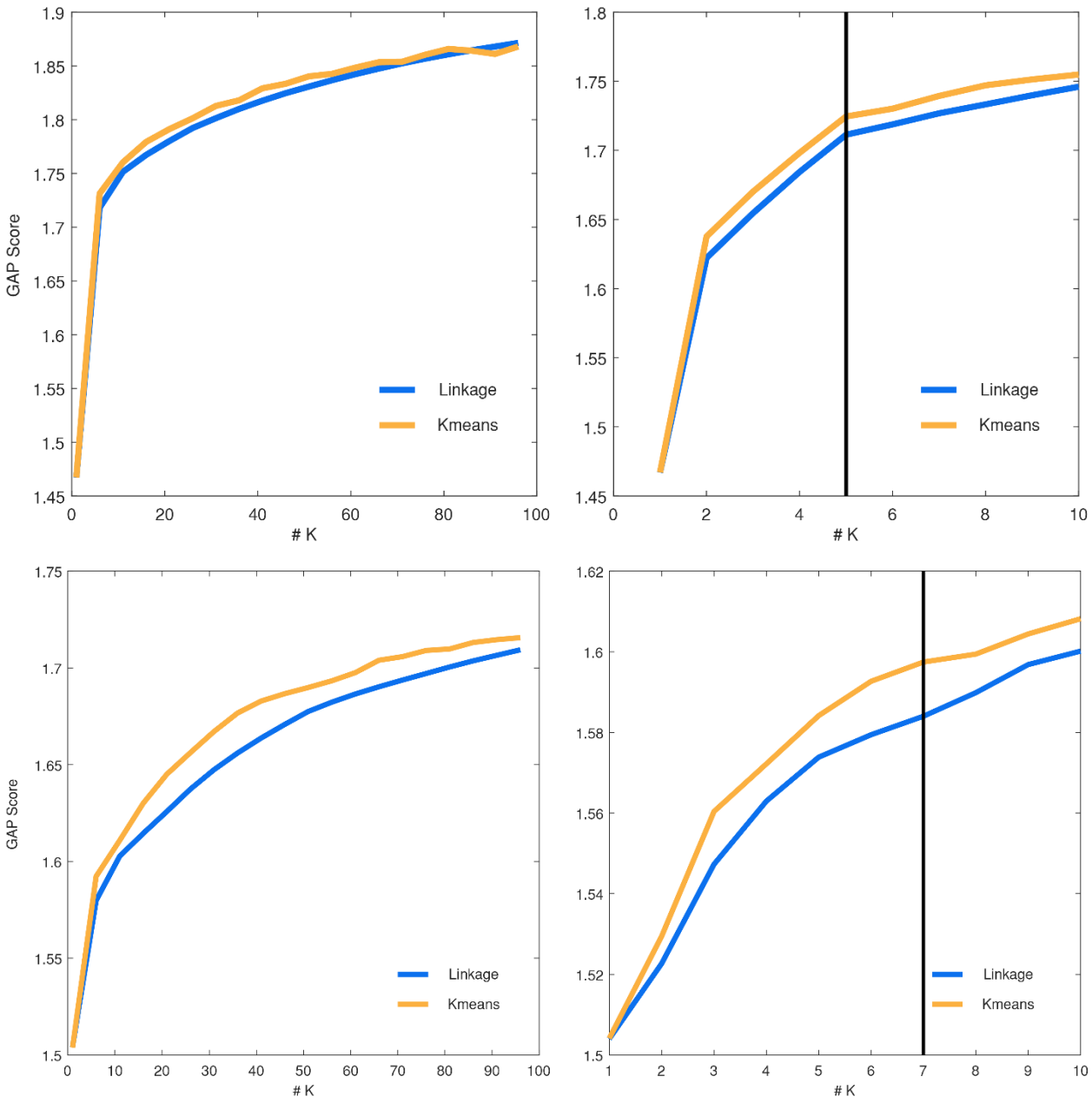



*Figure 83: Determination of optimal number of clusters using GAP method. GAP score was determined for various numbers of clusters (K) using both Linkage (Blue) and K-means (Yellow) clustering methods. This was done with the full data set (Top) and the data set reconstructed without the first two principle components (Bottom). The vertical black line represents the selected K value for each data set.*

# 4 Molecular Barcoding for Screening of Antibiotic Combinations

*With **Angga Perima***

A current limitation with combinatorial screening is the rapidly increasing sample number. This can make experiments with a large number of conditions difficult to screen with traditional microtitre plates due to the costs and time required. Droplet-based microfluidics addresses this problem by using picolitre droplets which act as individual reaction chambers [149]. This allows upwards of a million samples to be analysed within a single hour. However challenges remain in tracking the contents of each individual droplet through various microfluidic processes.

Here, we describe a method for associating specific environmental conditions (here antibiotic concentrations) to a unique barcode. Environmental conditions can then be randomly combined to observe their combinatorial effect on cell fitness. In terms of antibiotic resistance, this allows us to identify drugs that demonstrate a synergistic or antagonistic response when used in combination. This would have immediate implications in treatments for multiple-drug resistant bacteria. The barcodes we have developed are DNA oligomers, which when in combination become covalently linked by the DNA polymerase Klenow fragment. Droplets can then be split into separate populations by a selection criteria, which for our purposes is bacterial growth. These linked barcodes can be recovered from the droplets and sequenced to reconstruct the combination of antibiotics contained in each group of droplets.

We first demonstrate that this technique is compatible with droplet-based microfluidics. We then demonstrate that we can insert an artificial selection marker with specific barcodes and recover these markers with next-generation sequencing. This approach can be generalized beyond antibiotics to track any combination of environments effect on a fluorescence-linked selection criteria.

The design, characterization and optimization of the microfluidics for combinatorial antibiotic screening using droplet based microfluidics was performed by Angga Perima. Further details regarding microfluidic details are available in his doctoral thesis entitled Combinatorial Antibiotic Screening using Droplet Microfluidics (2017). Please note that the information in this chapter is confidential pending an ongoing patent application.

## 4.1 Tracking Droplet Combinations with DNA barcodes

Antibiotic resistance has become a severe issue to global health. Multi-drug resistant bacteria threaten to return us to a world where currently minor infections become life threatening. The WHO has identified pathogens in which resistance to antibiotics has become a critical threat, including common hospital acquired infections such as *Pseudomonas aeruginosa* and methicillin-resistant *Staphylococcus aureus* (MRSA) [150]. Despite this pending threat there have only been two new classes of antibiotics developed and approved in the last 20 years, lipopeptides and oxazolidionones. One approach to tackling this problem is discovering new antibiotics, perhaps through new methods of culturing previously unculturable bacteria. This has shown some success in finding new classes of antibiotics [151] [152] however discovery is only the initial barrier. The net present value, or sum of all investment costs and expected future revenues of antibacterial drugs is -42.61 million dollars [153]. This is because it can cost up to 2.6 billion dollars to develop a new drug [154] and any new antibiotic drugs which are developed are immediately restricted to drugs of last resort. This restricts the use and therefore market of new antibiotics specifically because of this very problem of spreading multiple drug resistance. An alternative approach involves systematically screening combinations of antibiotics for drugs with synergistic effects, or which can kill bacteria in combination that would otherwise survive either drug alone [155].

The high throughput nature of droplet-based microfluidics has recently been used to tackle the high sample numbers produced when approaching drug combination screening and overcome the cost and time limitations to screening using robotics and microtiter plates [156] [157]. However the difficulty in droplet based microfluidics is tracking the contents of each droplet. The approach employed by Eduati et al uses microfluidic valves to control the quantity of each compound for each droplet [157], this inherently limits the number of compounds that can be screened in a single experiment to the complexity of the microfluidic device. Additionally, the order in which droplets are produced must be maintained to identify droplets with their contents, and thus they must be incubated in long microfluidic tubing to maintain the correct order. The approach taken by Kulesa et al on the other hand associates each antibiotic to a fluorescent dye [156]. While this simplifies handling, the method of droplet pairing and fusion is again limited in through-put by the design of the microfluidic chip, and the limited number of well separated fluorophores again limits the number of compounds that can be used in a single experiment. An alternative to fluorescent barcodes are molecular barcodes, which take advantage of modern high-throughput sequencing techniques to encode a specific barcode onto a DNA oligomer.

Molecular barcoding allows us to associate a condition, in this case a specific antibiotic concentration, with a short sequence of DNA. This approach has several advantages over optical barcoding. First, fluorophores are prone to optical leakage, and as such the peak excitation and emissions for optical dyes must be well separated to prevent false detection. This limits the number of samples that can be analyzed at a time. In contrast, DNA barcodes allows for $4^n$ unique sequence combinations, where n is the number of nucleotide bases in the barcode. As such even a short 10 nucleotide barcode allow for tracking of potentially a million unique conditions. Secondly, DNA barcodes are compatible with optical labelling, allowing for optical dyes to be used to quantify and manipulate droplets such as with fluorescent activated droplet sorting (FADS) [149]. And finally, the cost to synthesize DNA barcodes is inexpensive, at only 0.25 EUR per base for 250nmol, with the cost of synthesizing and sequencing DNA regularly beating Moore's Law with the cost of sequencing per raw megabase of DNA decreasing by $10^5$ from 2001 to 2015 [158].

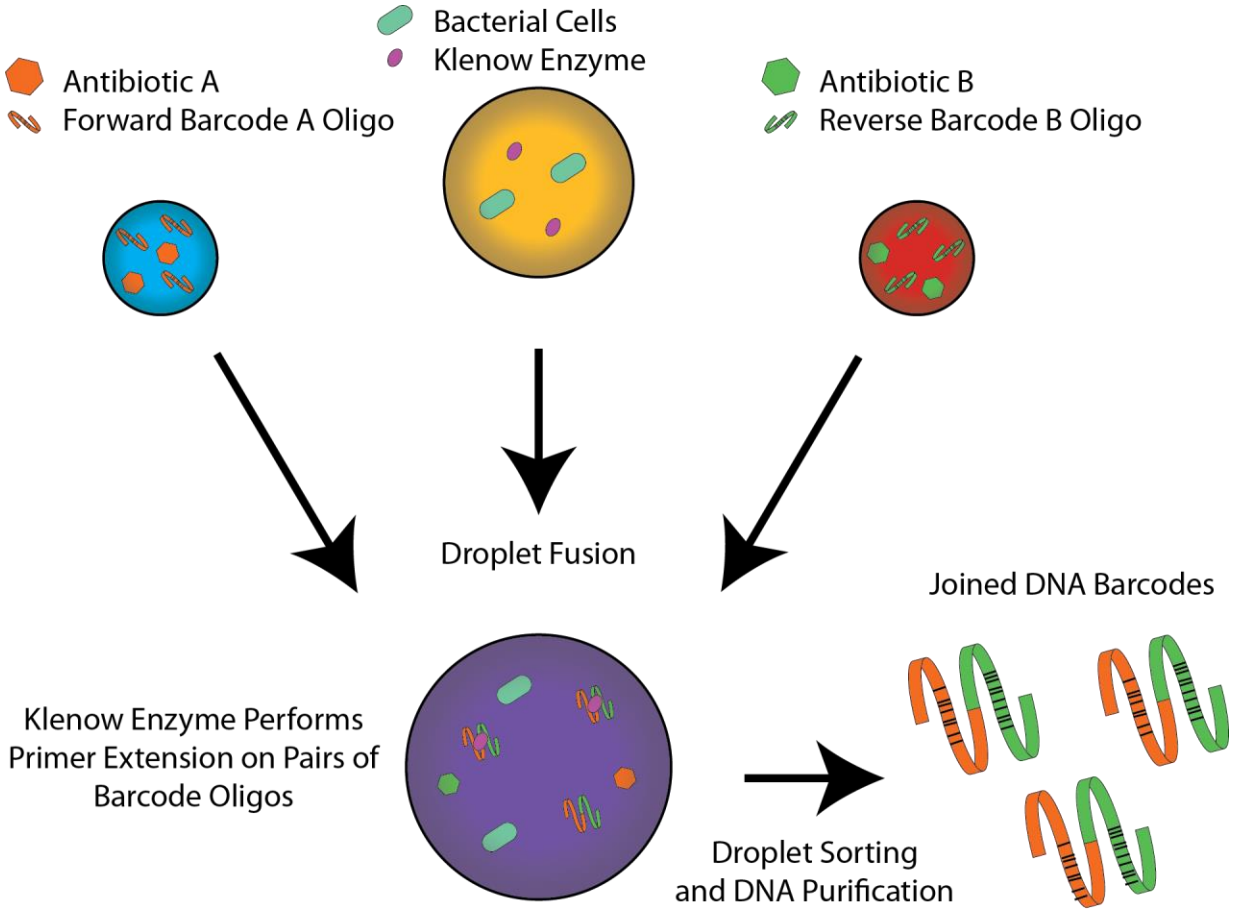



*Figure 84: Schematic of molecular barcoding technique for tracking antibiotic combinations in picolitre droplets. Antibiotics concentrations can be tracked by associating them with a short (<60 nt) DNA oligo that contains a 16 nt barcode unique for that antibiotic concentration. This is repeated for both a forward and reverse set of barcoded oligos for each antibiotic and range of concentrations desired. The antibiotic/barcode mixtures are then encapsulated into ~10 pL droplets. Two of these droplets are then paired with an ~40 pL droplet containing bacterial cells and Klenow enzyme. When the two droplets containing antibiotics / barcode oligos are complementary forward / reverse, they will anneal and the Klenow enzyme will covalently link them through primer extension. Droplets can then be sorted by bacterial growth, and the barcodes recovered by DNA purification. Only correctly paired barcodes will be sequenceable (approximately 50% with random pairing). This allows detection of which combinations of antibiotics (and their doses) allow bacterial growth, and can be generalized to track the effect of any combination of environments (here antibiotic concentrations) on any selection that can be coupled to fluorescent detection (here bacterial growth).*

Our molecular barcoding strategy involves directly associating DNA oligos to specific concentrations of antibiotics. We currently have 96 unique forward barcodes and 96 unique reverse barcodes. This allows us to test up to 96 antibiotic concentrations in a single experiment. Solutions are prepared in a microtitre plate containing the barcode and antibiotic, which is used with a 96 parallel droplet maker to create an emulsion where each droplet contains one DNA barcode and one antibiotic. This is then repeated with the reverse barcodes. Next, droplets containing bacterial cells and Klenow enzyme are formed, and these droplets are fused with 1-2 droplets containing either forward or reverse barcodes. In the droplets with one forward barcoded oligo and one reverse barcoded oligo, a complementary linker region will allow the two oligos to anneal, and the Klenow enzyme will perform primer extension to fill in the remaining sequence, covalently linking the two barcoded oligos. As such, only droplets with at least one forward and one reverse barcoded oligo will be suitable for amplification and sequencing, and droplets with only one oligo or two oligos in the same orientation will have no reaction. Droplets can then be sorted by the

number of bacteria in each droplet, and droplets where bacteria were able to grow in the presence of antibiotics can be sequenced separately from droplets in which bacteria were unable to grow. DNA oligos from each pool of droplets can then be isolated, purified, amplified and sequenced to determine the antibiotic combinations that resulted in each phenotype.

Our DNA barcodes are designed to follow specific criteria. They are all exactly 16 nucleotides long. This allows each barcode to be well separated in sequence to prevent any PCR errors or sequencing errors from changing one barcode to another. Barcodes are made by a python script in which a random number generator is used to determine a random 16 nucleotide sequence. This sequence is tested to ensure that it has at least 3 nucleotide differences between any previously generated barcodes. The linker sequence and the illumina primer sequences (Rd1 with forward barcodes and Rd2 with reverse) are then joined to the barcode sequence and the oligo is checked to ensure that it does not form any false priming sites, hairpins, primer-dimers, and have a similar %GC content. If any of these criteria are not met, the barcode sequence is rejected. This was done to create 96 unique forward barcodes and 96 unique reverse barcodes, although this can continue to be scaled up if required. When two DNA barcodes are linked, the resulting sequence is 96 nucleotides long, and contains the two DNA barcodes connected by the linker, and flanked by the Rd1 and Rd2 illumina sequences. The Rd1 and Rd2 sequences can then be used as primers to add the P5 and P7 sequences, as well as any illumina indexes desired to further multiplex experiments.

## 4.2 Quantification of barcode efficiency

***Fluorescent dyes, growth media, and encapsulation do not inhibit Klenow reactions***

We first tested that our barcodes were able to efficiently be linked by primer extension with Klenow enzyme. We did this by mixing forward and reverse barcodes in a klenow reaction for one hour. We then purified the reaction and amplified the product with PCR using primers containing 5' extensions of the P5 and P7 illumina sequences. This PCR product was again purified and the product was amplified finally with P5 and P7 illumina primers. We then ran each product (including the annealed but not extended initial DNA oligos) on an agarose gel to visual the DNA. As expected, the size of the DNA band corresponded to what was expected, and the size of the DNA extending with the subsequent PCRs. Additionally, we sequenced the DNA product, which matched the expected sequence. We repeated the same protocol, however with the Klenow Reaction performed in droplets with the addition of growth media, bacteria and antibiotics. We found the same expected products as in the earlier bulk test.

To ensure that the fluorescent dyes used in the microfluidics were compatible with the Klenow reaction, we tested the Klenow reaction with a variety of fluorescent dyes. After purifying the Klenow reaction, samples were amplified by qPCR to quantify the quantity of template produced in each reaction. We found no noticeable difference in qPCR quantity between any of the dyes and the control without any dye (Figure 85). We also tested to see if the addition of growth media effected the Klenow reaction due to the extra salts introduced. We added LB media, M9 Media, and M63 Media to Klenow reactions, as well as testing Klenow reactions with the aforementioned media replacing Klenow buffer rather than supplementing it (with an additional dosage of $MgCl_2$). We found that none of the media reduced the Klenow efficiency when supplementing Klenow reactions, however the Klenow buffer was necessary for efficient reactions to occur as $MgCl_2$ was not sufficient to replace it. Finally, we confirmed that confinement in droplets was not abolishing Klenow efficiency by creating a solution with all the components expected in our experiments final droplets, at the expected concentrations. We then used some of the solution to create droplets, and the remaining solution was incubated in bulk. We also made the same solution without any Klenow enzyme as a negative control. We found that the droplet solution had a reduced efficiency than the bulk solution, however we still had a strong detectable product.

***Mungbean Endonuclease prevents unused barcodes from contaminating downstream PCR***

Once the DNA barcodes have been joined by primer extension, they are purified from non-joined oligos using Mung Bean Nuclease. This digests only single stranded DNA. Any oligos that did not undergo primer extension are destroyed. Unused oligos are further reduced from size exclusion during DNA purification, where the exclusion limit is 100 bp, and individual oligos are less than 60 nucleotides. It is important that unused oligos are removed to prevent them from acting as primers in the subsequent downstream PCR reactions. This would result in incorrect barcode combinations. We tested the frequency of mismatching barcodes by performing separate Klenow Reactions for different pairs of barcodes. Each pair of barcodes were purified with Mung Bean Nuclease, then pooled and amplified by PCR. We sequenced the amplified DNA and tested how many reads contained the correct combinations and how many contained combinations between barcodes in different Klenow Reactions. We found that we had 96% and 98% of reads containing with the correct combination of the first and second forward barcodes paired with the first and second reverse barcodes respectively. We found only 0.2% of the incorrect combination of the first forward barcode paired with the second reverse barcode and no incidents of the

second forward barcode pairing with the first reverse barcode. One of the two barcodes within remaining reads we were unable to be successfully mapped.

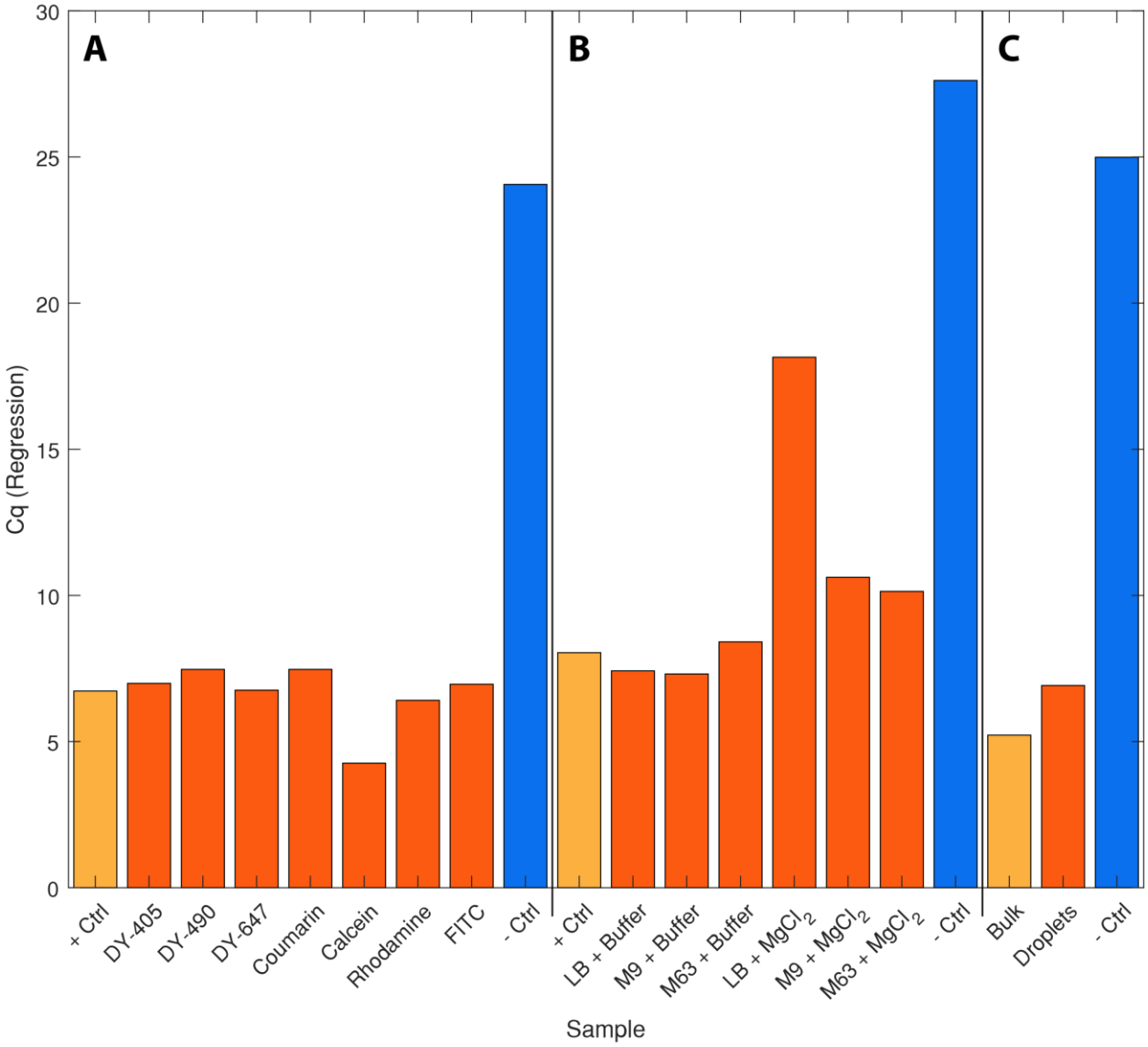



*Figure 85: Quantification of the impact of various additives to primer extension efficiency with Klenow by qPCR. We tested the impact of fluorescent marker dyes (A), growth media (B), and droplet encapsulation (C) on Klenow efficiency. Yellow represents a standard Klenow reaction as a positive control, blue represents a negative control with no Klenow enzyme and orange represent test samples. Dyes were added to standard Klenow reactions at 5x working concentrations. Media was used to replace water in a standard Klenow reaction, with 1x Klenow buffer or only 1x MgCl2. Droplet encapsulation was done with the same solution used for both a bulk incubation and incubation done within ~60 pL droplets. This solution had the equivalent concentrations of a true experiment for all necessary reagents.*

***Specific antibiotic concentrations can be tracked through droplet microfluidics with DNA barcodes***

We have performed a cell-free proof of concept experiment to demonstrate the technology, in which selection is dependent on the presence of a selection dye rather than the concentration of bacteria. Here, we used a reduced number of barcodes (48 forward and 48 reverse). In which half of the reverse barcodes are tagged for positive selection (Figure 86). This demonstrates that any assay that can be linked

to a fluorescent readout (and thus can be sorted with FADS), would be compatible with our system. We created 10 pL droplets with a 96 parallel droplet maker containing barcoded oligos and merged them with 40 pL leader droplets containing Klenow enzyme. All of the 10 pL droplets contained the fluorescent dye DY-405, which allows us to quantify how many 10 pL droplets fuse with each 40 pL droplet. Half of the reverse barcoded oligo droplets also contained dye DY-647 which was our selection criteria. We detected fluorescence in each droplet for each of the corresponding dyes, and when plotting the distribution of each of these signals before droplets were sorted, we can see clouds of droplets that correspond to fused droplets containing 1, 2, or 3 oligo containing droplets, and 1, or 2 oligo droplets containing selection dye. We sort droplets containing 2 oligo droplets, 1 of which contains the selection dye. The rest of the droplets were unsorted. After positive sorting of >40,000 droplets, we broke the emulsion of the sorted and unsorted droplet populations, as well as the remaining droplets that had not been used to represent the initial droplet population. These three samples were purified as described above and amplified by PCR to add the P5 and P7 regions, as well as an illumina index for each population to demultiplex them in the sequencing data. We only recovered a small quantity of DNA barcodes after purification but we were able to see a clear enrichment of the intended barcodes in the sorted droplet population. We noticed the dNTP

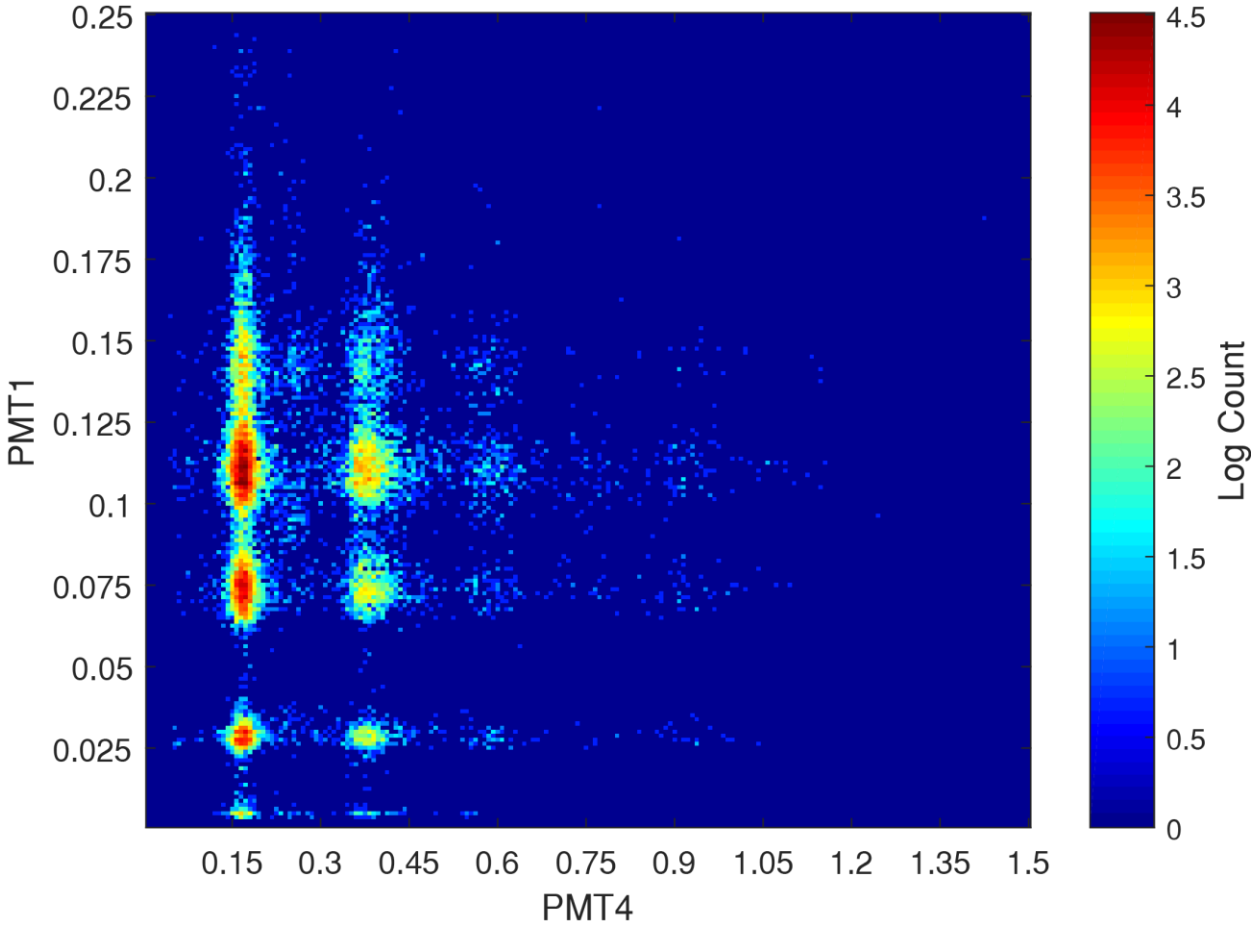



*Figure 86: Droplet populations for fluorescence activated droplet sorting. PMT1 corresponds to droplets containing 0, 1, 2, or 3 small barcode containing droplets (from bottom to top). Populations from left to right (PMT4) corresponds to droplets containing 0, 1, or 2, selection criteria. The desired droplet population is selected as the droplet population corresponding to 2 small droplet barcode containing droplets and 1 selection criteria droplets (PMT1 0.1-0.125, PMT4 0.3-0.5).*

concentration used in our solution was insufficient for primer extension with all of our oligos within a droplet. Increasing the dNTP concentration resulted in 15 fewer cycles required during PCR for sufficient product for sequencing. It is currently pending sequencing.

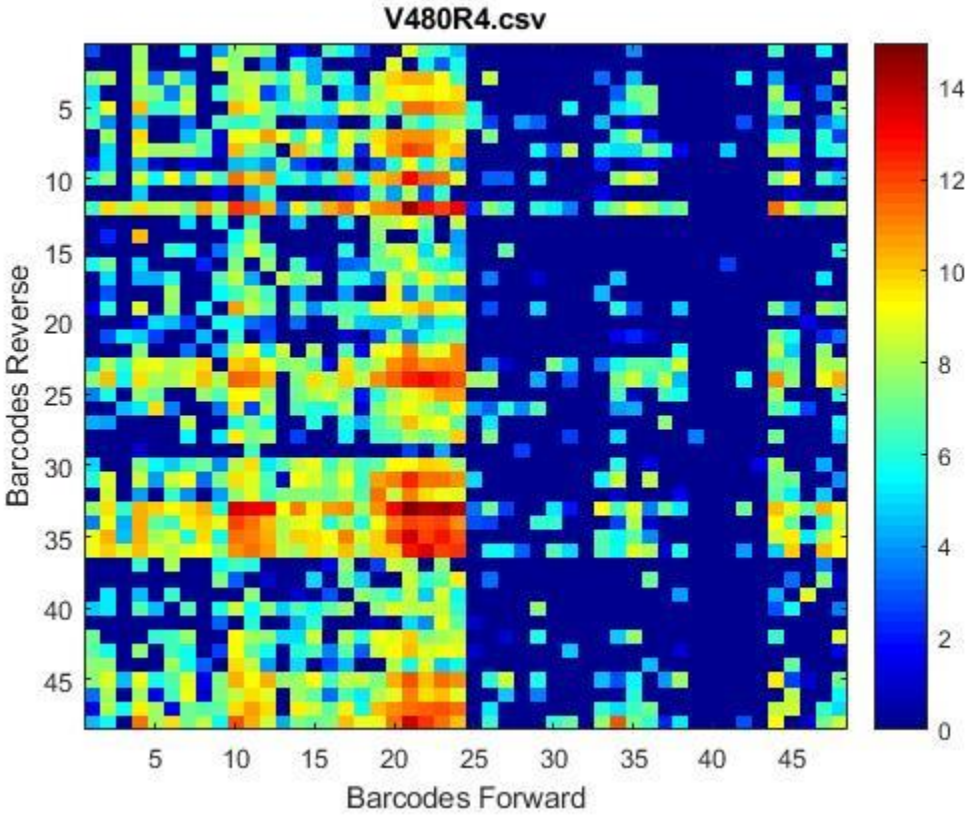



*Figure 87: Sequencing reads for each barcode pair in a logarithmic scale. The left-hand side was spiked with dye for selection with fluorescence activated droplet sorting.*

This method of molecular barcoding provides several advantages over the methods previously described using microfluidic valves [157] or optical barcoding [156]. Multiple 96-well parallel droplet makers can be used in a single experiment to increase throughput. This limits the number of compounds which can be tested in a single experiment to those which are compatible with fluorinated oil and droplet microfluidics in general, as the number of unique barcodes can easily be expanded for additional samples. Secondly, it is possible to combine more than 2 conditions with DNA barcodes, as primer extension could be done with greater than 2 DNA oligos. Thirdly, there is no limit on the number of droplets that can be processed in a single experiment, and this is currently only determined by the frequency of droplet sorting and the amount of time dedicated to each step, rather than any design features. Finally, as each part of our microfluidic pipe-line is modular, effectiveness improves with new designs for droplet generation, droplet fusion-pairing, and droplet sorting. The design of the molecular barcodes is compatible with multiple approaches for each of these.

## 4.3 Methodology

*Barcode design*

Barcodes were created with a custom python script. A random number generator is used to create a random sequence of 16 nucleotides. A 20 nucleotide linker region and a 20/22 nucleotide region of either Rd1 or Rd2 was then added to the random barcode sequence. Primer3 binaries were then used to quantify any primer dimers or hairpins, and the Tm and %GC content of the 16 nt barcode sequence. Barcodes that passed all selection criteria were then checked for Levenshtein distance against already existing barcodes. As long as they had a distance of > 3 with all other barcodes, the barcodes were accepted. Barcodes were synthesized in 96 wells at 6 nmol scale by IDT. We added 1 mL of nuclease-free water to each oligo to rehydrate them at a concentration of 6 mM.

*qPCR quantification*

Klenow control reactions were done with 2 µL Klenow buffer, 1 µL dNTPs (10 mM), 1 µL Klenow enzyme, 1 µL Forward Barcode (6 mM), 1 µL Reverse Barcode (6 mM), 14 µL nuclease-free water. To quantify the effect of fluorescent dyes, 4 µL of Dye (100mM) was added to the Klenow reaction (reducing water to 10 µL). To quantify the effect of growth media, nuclease-free water was replaced with either LB, M9 or M63 media. To quantify the effect of droplet encapsulation, a solution containing 5 µL of Forward Barcode, 5 µL Reverse Barcode, 5 µL dNTPs (10 mM), 5 µL Klenow Enzyme, 10 µL of Dye-405, 10 µL Klenow Enzyme and 60 µL of LB media was made, split in 2 and kept on ice. Half was used to make ~60 pL droplets. Droplets and the remaining solution were incubated at 37 $^{o}$C for 1 hour. qPCR was performed using Agilent Brilliant III qPCR master mix.

*Droplet-Based Primer extension*

96 wells within a 384 well microtiter plate were filled with 5 µL Dye-405 (100 mM), 1 µL either Forward or Reverse Barcode (6 mM), 10 µL dNTP (10 mM), 5 µL Klenow Buffer, 5 µL of Antibiotic (concentration dependant on MIC) and 24 µL of LB media. This plate was used to create ~10 pL droplets using a 'hedgehog' design previously described by Angga Perima. A solution containing 20 µL Dye-647 (1 mM), 2 µL E. coli culture grown overnight in LB, 10 µL Klenow Enzyme, 20 µL Klenow Buffer, 40 µL dNTP (10 mM), and 108 µL of LB media was made and used to create ~40 pL droplets. Large 40 pL droplets were merged with smaller 10 pL droplets at a ratio of 1:2. Droplets were inclubated at 37 $^{o}$C for 3 hours and then pico injected with syto9 dye. Droplets were sorted by their fluorescent from Dye-405 (the number of small droplets), and Syto9 (the number of bacteria). Emulsions were broken by removing excess oil and adding equal volume of Perfluoro-Octanol. Details regarding microfluidics can be found in the thesis of Angga Perima.

*Barcode Clean-Up and Purification*

DNA was extracted from droplets (or bulk Klenow reactions) using a Macherey-Nagel PCR Clean-up kit, with 200 µL of NTI buffer and 100 µL of Barcode solution (solutions increased to 100 µL if not already). This results in size exclusion cut off of ~100 bp. DNA was eluted in 44 µL of water and added to 5 µL mung bean endonuclease buffer, and 1 µL of mung bean endonuclease. Solution was incubated for 30 minutes to destroy any non-extended barcodes. Solution was increased to a volume of 100 µL and purified again with a Macherey-Nagel PCR Clean-up kit.

# 5 Single Cell Transcriptional Analysis of *E. coli*

Droplet-based single cell RNA sequencing has recently been developed for mammalian cells [17] [18]. This allows for the transcriptional analysis of thousands of individual cells in parallel. As a result, phenotypically distinct (though potentially genetically identical) subpopulations can be identified within a sample. In traditional RNA sequencing or microarray techniques the distinct transcriptional profiles of these subpopulations would be lost by being combined together. This also allows for the potential to screen high-throughput perturbation libraries such as those generated with CRISPR-Cas [20].

Here we describe adapting droplet-based single cell RNA sequencing to bacterial samples. This would allow us to greatly increase the number of genes that we could perturb to cover all global transcription factors, and potentially all sigma factors as well. However there are key challenges in working with bacterial cells. These include the bacterial cell wall, which makes cell lysis more difficult; the lack of poly-A tails on mRNA, which require using gene specific primers for the reverse transcriptase; and the contamination of genomic DNA, which appears identical to cDNA due to a lack of introns and poly-adenylylation. While all of these are easily addressed in bulk RNA-sequencing, with droplets we are unable to purify samples after encapsulation until all the RNA has been tagged with a droplet specific barcode. This means that cell lysis, reverse transcription, and removal of genomic DNA must all be compatible.

After successfully sequencing RNA from single bacteria, we found that the transcriptional data is very sparse, even more so than the notoriously noisy single cell data from mammalian cells [110]. Finally we discuss a collaboration with Institut Pasteur to apply this technique to the study of antibiotic persister cells, a distinct subpopulation in bacteria cultures which are able to survive short-term exposure to antibiotics despite not carrying any genetic antibiotic resistance.

## 5.1 Microfluidic Design for Single Cell Transcriptional Analysis

Recently there has been considerable development of single cell analysis and sequencing technologies [17] [18] [107] [159]. Single cell level analysis opens up the possibility to disentangle complex biological samples and analyze large sample libraries while remaining cost efficient. Microfluidics plays a central role in these developments as it allows to highly parallelize the manipulation of cellular material in ~10 to 100 pL compartments, which function as the equivalent of miniaturized reaction wells. The scale of the droplets is ideally suited for single-cell analysis. Two main microfluidic approaches have prevailed. The first relies on miniaturized hydro-pneumatic valves, allowing sequences of fluidic operations to be performed in micro chambers. This technology led to the first commercial microfluidic system for single cell sequencing, the C1 from Fluidigm, which allows the analysis of at most several hundred cells [160]. However, despite recent improvements, cell loading remains unreliable and even impossible for many cells types including bacteria, as they do not fit within a specific range of sizes, morphologies and adhesion properties. Given these limitations, single-cell protocols in multi-well plates, where the number of cells that can be sequenced is similar, tend to be favored in practice [161]. The second approach is droplet-based microfluidics, combined with molecular barcoding. In droplet-based microfluidic systems pL to nL volume droplets in an inert carrier oil are used as independent micro reactors compartmentalizing single cells. Here, throughput can easily reach millions of single cells per hour, and is currently only limited by downstream sequencing capacities. Three commercial droplet-based microfluidics systems for single-cell RNA-seq (3'-end sequencing of poly(A) mRNA) are now available. In 2016, 10x Genomics launched the Chromium System and 1CellBio introduced the inDrop system, both capable of analysing up to 48,000 cells per experiment. In 2017, Illumina and Bio-Rad launched the Illumina Bio-Rad Single-Cell Sequencing Solution, capable of analyzing from 100 to 10,000 cells per experiment. However, there currently does not exist any technological option for RNA-seq of single pathogenic organisms, including bacteria, viruses and unicellular eukaryotic parasites. A number of challenges need to be overcome. Taking the example of transcriptomic analysis of bacteria, a droplet-based microfluidics system would need to:

1. lyze bacteria in droplets
2. perform RT in droplets on mRNA lacking polyA tails;
3. remove background signal from genomic DNA (as mRNA lacks poly(A) tails and introns);

Overcoming these challenges requires the molecular biology of DropSeq [17] [18] to be significantly altered. We use a PEG/Acrylamide Hydrogel Bead to carry our molecular barcodes, which have been redesigned to be optimized for use with Eubacteria, Archaea, Human, Mouse and Many model organisms. Each barcode is 16nt long with a minimum annealing temperature of 40°C, there are at minimum 3 degrees of error between all barcodes, and barcodes are blasted [162] against the Representative Genomes Database (containing Eubacteria and Archaea genomes), ESS database, and the Human and Mouse Transcriptome databases to reduce the probability that they will act as false primers. These barcodes are ligated onto our Hydrogel beads using a split/pool method (Figure 88A) in which the beads are equally distributed among a microtitre plate of 96 barcodes. All of the ligation sites on a given bead then receive exactly the same barcode. After ligation of one index of barcodes, the beads are pooled and washed to remove any non-ligated barcodes. We then repeat the bead distribution and ligation with the second, third, and fourth indexes of barcodes. This strategy ensures that all primers on a given hydrogel bead have exactly the same primer, yet there are 96^4 (~85 million) possible paths for each

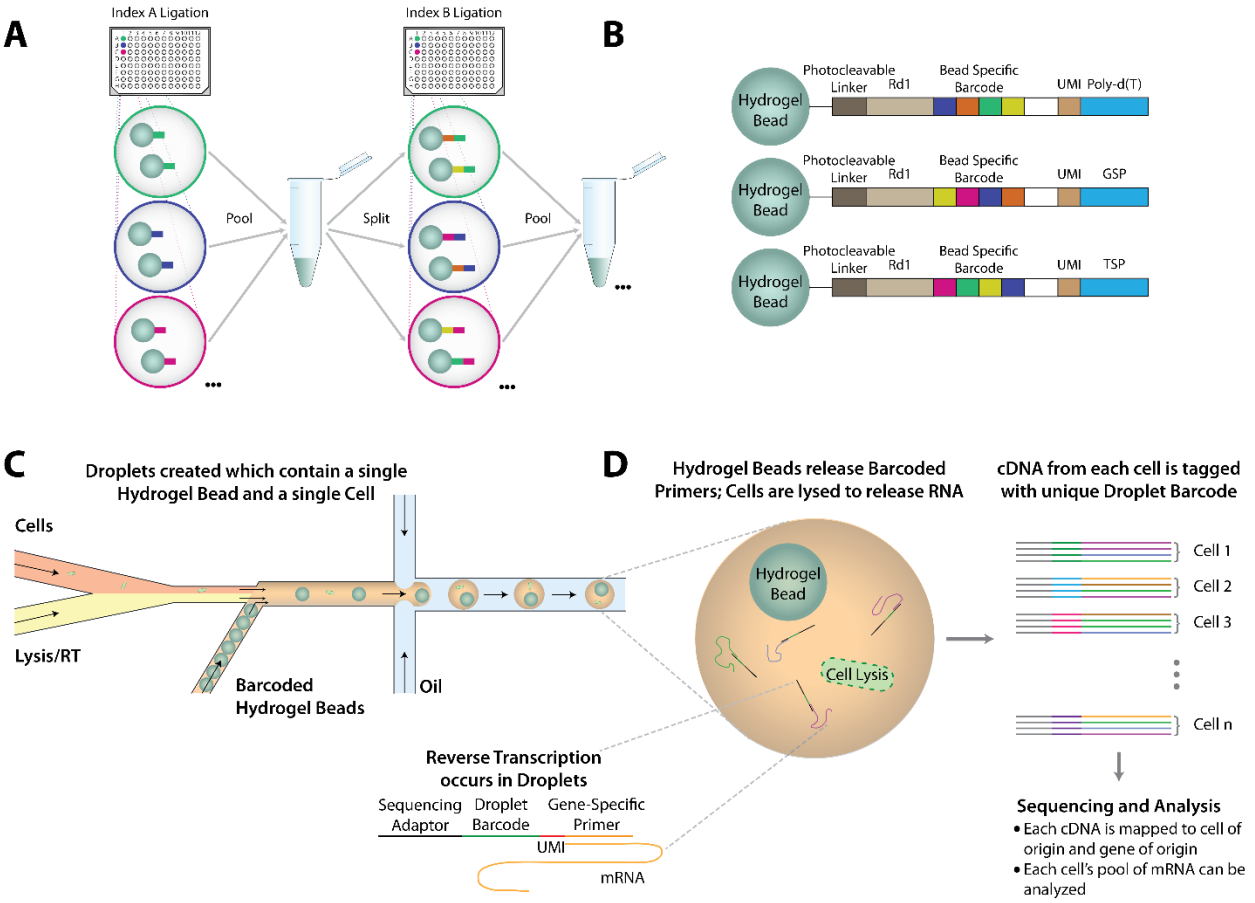



*Figure 88: Overview of single cell RNA-seq with Droplet based Microfluidics. A) Split-and-pool synthesis of barcode libraries. Each 16 bp DNA index is represented by a color. Polyethylene Diacrylate (PEG-DA) hydrogel beads (HgBs) decorated with double-stranded DNAs coupled to the beads via a 5' acrylate are distributed into wells of a 96-well microtitre plate, each containing a different first index (index A) and index A is added by ligation. The HgBs are then pooled, washed and re-distributed into the wells of a second microtitre plate, each containing a different Index B, which is ligated to Index A. Repeating this splitting and pooling process 3 times in total (adding 3 indexes) results in 963 combinations, which generates ~106 different barcodes. If further diversity is required, a fourth index can be added resulting in 964 combinations, which generates ~108 different barcodes. B) After adding the last index, the beads are pooled once again, and a cocktail of DNA molecules, each containing a unique molecular identifier (UMI) sequence and primer regions are ligated to the barcodes on the beads. This primer region depends on the application; either a polyd(T)VN for 3'-end total RNA-seq, or a mix of Gene Specific Primers (GSPs) for targeted RNA-seq. The second strand of the primer is then removed by alkaline treatment and a short DNA oligo complementary to the region containing the restriction re-annealed. Each HgB ends up with a total of 109 primers carrying the same bead-specific barcode. C) Microfluidic co-compartmentalization of single cells and single barcoded hydrogel beads in droplets with RT and lysis reagents. D) Process in droplets. Cells are lysed to release RNA and barcoded primers are released from hydrogel beads by UV cleavage of a photosensitive linker to prime cDNA synthesis. As each bead carries primers with a unique barcode, the cDNAs from each droplet carry a unique barcode, allowing them to be identified after sequencing.*

hydrogel bead to take through all four indexes; the barcode for every hydrogel bead is unique. Once the Hydrogel beads are barcoded, we can attach an adapter to them (Figure 88B). In the case of current single cell RNA sequencing with mammalian cells, a poly d(T) primer is used, however this is not suitable for bacterial cells so we replace it with a gene specific primer. Once our Hydrogel beads are complete, we encapsulate single bacteria cells into droplets along with hydrogel beads bearing our primers, and lysis and reverse transcription reagents (Figure 88C). This is done in a flow focusing droplet maker, where the three aqueous phases (containing cells, RT/Lysis, and Hydrogel beads) mix before the oil phase pinches off droplets. The deformable nature of the hydrogel beads means that they are able to beat Poisson

distribution and one hydrogel bead is contained in each droplet. Reverse transcription then occurs within droplets containing cells (Figure 88D) with primers released from the hydrogel bead. As a result, all the cDNA produced from a single cell will have exactly the same DNA barcode incorporated. Thus when we break our emulsion and proceed with sequencing, we can trace each cDNA back to its cell of origin.

***Bacteria encapsulation in microfluidic droplets follows a Poisson distribution***

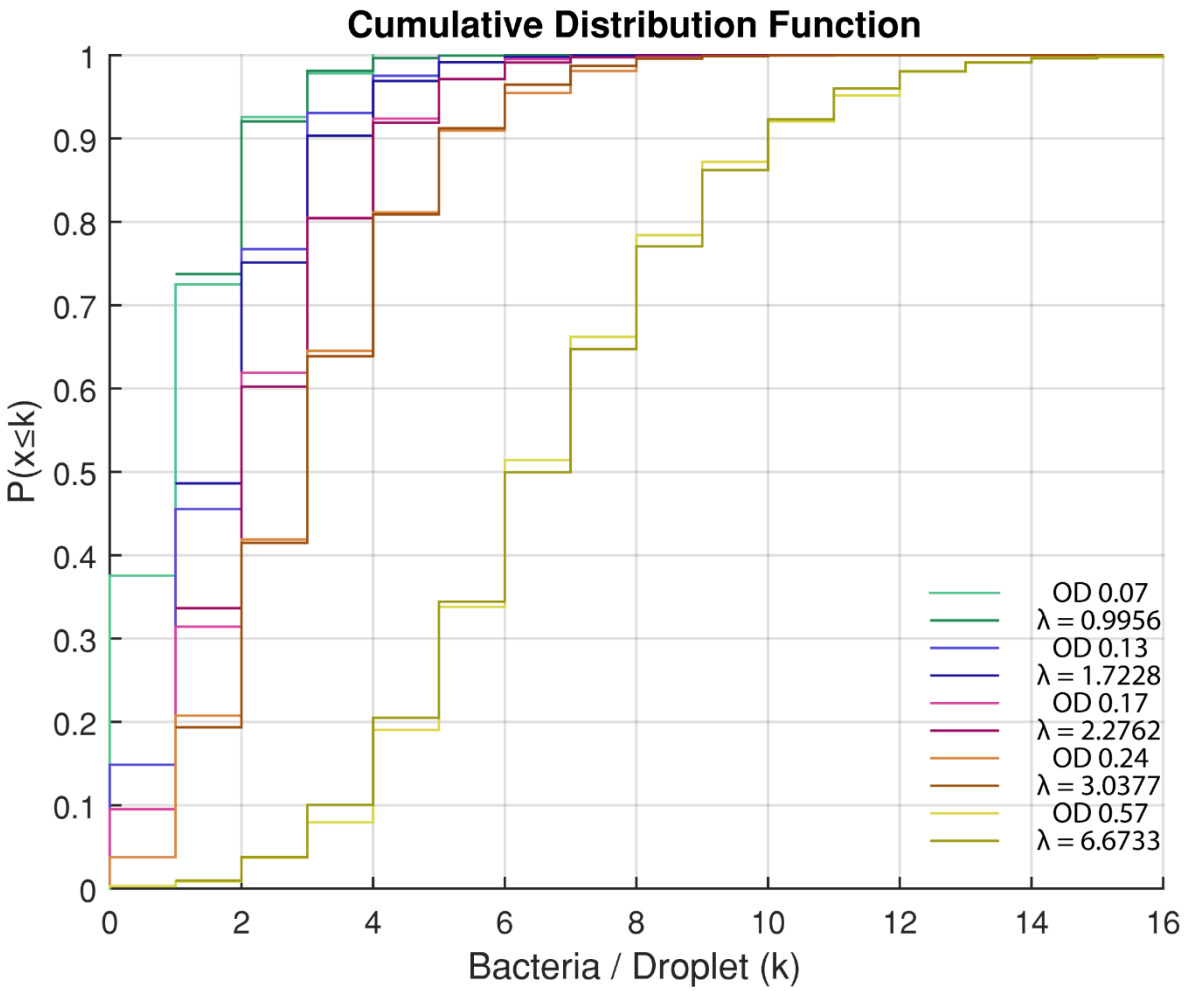



*Figure 89: Cumulative Poisson Distribution Function of Bacteria in droplets in bDrop-Seq protocol. The cumulative distribution function for the number of Bacteria in each droplet at 5 different ODs is shown with the corresponding lambda from a Poisson distribution. The encapsulation of MG1655 E. coli cells follows a Poisson distribution similar to that reported for mammalian cells, with an initial OD of 0.07 corresponding to a lambda of ~1.*

We tested if bacteria cells are encapsulated into droplets following the same Poisson distribution as mammalian cells. We did this by counting the number of bacterial cells per droplet at different starting ODs (Figure 89). This is important because it is not obvious that bacterial cells will not adhere to each other or the microfluidic chips walls. We also use this data to create a calibration curve (Figure 90) to determine what OD we should use to have a given lambda. We need a lambda of ~0.1 to ensure that cells we process are single cells and not encapsulated with additional cells.

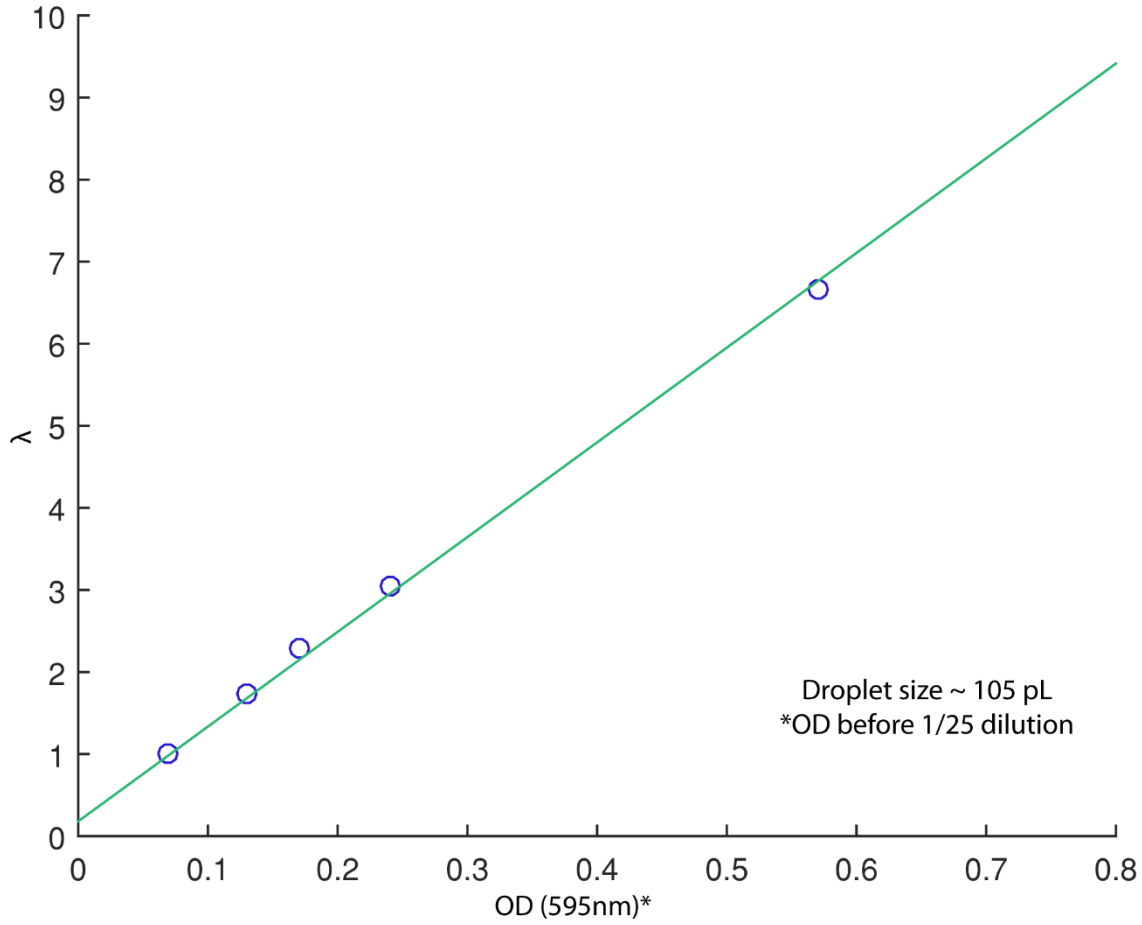



*Figure 90: Calibration curve for determining necessary OD for a desired lambda. The OD of samples prior to processing for bDrop-Seq is plotted against the resulting lambda for bacteria per droplet. To ensure that droplets contain at most 1 bacteria per droplet, a lambda of 0.1 is desired, corresponding to an OD of ~0.02.*

***RNA-Later but not RNA-Protect is suitable for single-cell droplet-based RNA-sequencing***

DropSeq for mammalian cells currently published recommends that any cell types used should have stable mRNA levels for at least 30 minutes so that mRNA is not degraded prior to reverse transcription in droplets. Since the half-life of mRNA in bacteria is significantly shorter at approximately 4 minutes, we take steps to stabilize the mRNA before the experiment proceeds by treatment with RNALater (Figure 91). We tested both RNAProtect and RNALater as mRNA stabilization reagents, however we found that RNAProtect was resulting in premature cell lysis and thus was not suitable for single cell analysis.

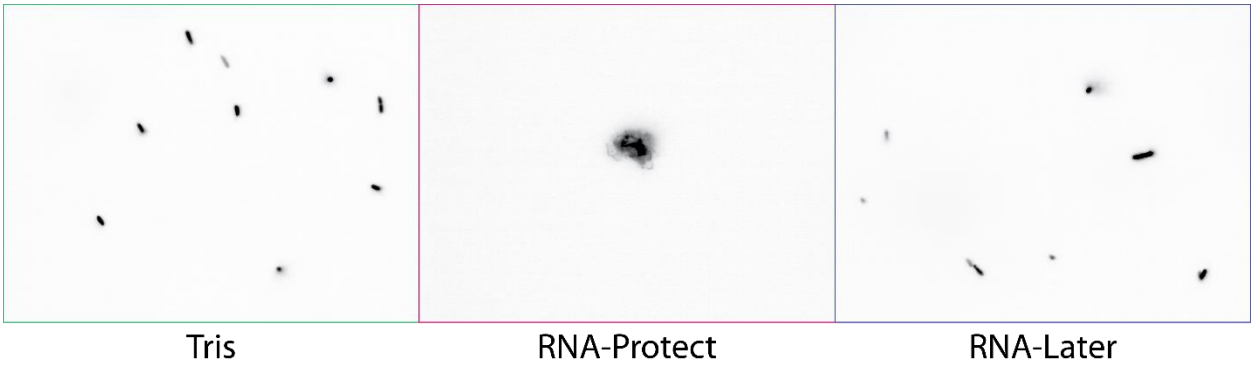



*Figure 91: Effect of different RNA-stabilization reagents on cell integrity. E. coli cells expression EGFP were viewed under a fluorescent microscope after being treated with Tris solution, RNA-Protect from Qiagen, or RNA-Later from Ambion. Cells remained intact with RNA-Later but not with RNA-Protect.*

***Microfluidic device for single cell droplet-based RNA-sequencing is a 3 inlet flow focusing device***

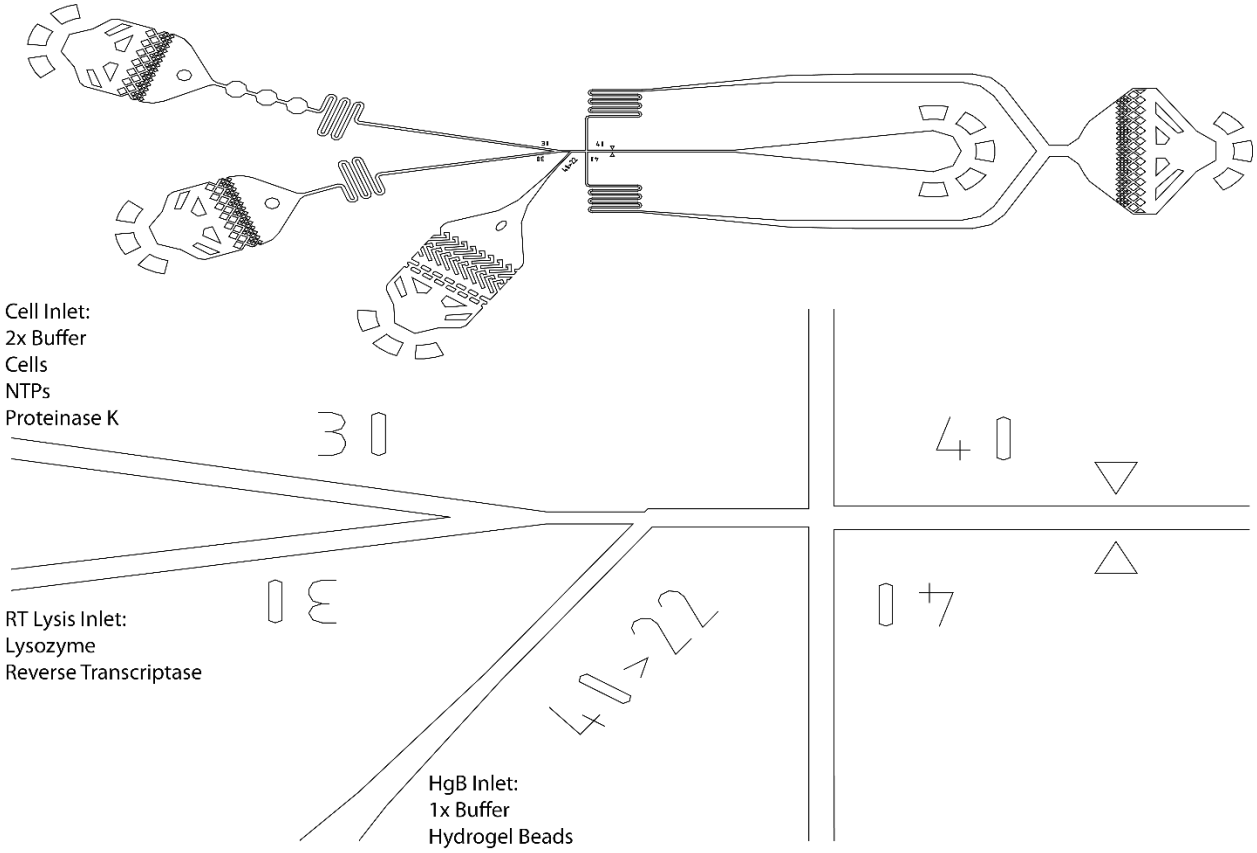



*Figure 92: Microfluidic Device for Bacterial single cell RNA-seq. The microfluidic device contains 3 aqueous phase inlets for the cell solution, the RT/Lysis solution, and the Hydrogel Beads (HgB). These aqeous phases co-flow until they intersect with a flow focusing droplet maker, which are then collected for incubation off chip.*

Finally, we had to modify the microfluidic chip design (Figure 92). Specifically, we noticed lysis was possible immediately upon contact between the lysis solution and cells within the cell solution, so the distance from when these two inlets create a co-flow to droplet formation was greatly reduced. Additionally, we added a notch at the hydrogel bead inlet to try to limit back-flow which could occur with polydispersity in the size of the hydrogel beads.

## 5.2 Cell lysis and Isolation of cDNA

***Bacterial cells are lysed with Lysozyme and Proteinase K***

The lysis of bacteria cells in droplets is significantly more difficult than mammalian cells due to the bacterial cell wall. Currently, we lyse bacteria cells with a combination of lysozyme and polymyxin B. A challenge with this approach is that lysozyme is sensitive to cation concentration and pH [163], preferring solutions with a low ionic strength and a low pH. This is in contrast to reverse transcriptase which prefers a high ionic strength and a high pH. In particular magnesium, which is required for reverse transcription, denatures and precipitates lysozyme. As such we have spent a significant amount of time creating custom buffers that allow lysis and reverse transcription to occur in the same reaction (Figure 93). Critically, lysozyme is kept separated from salts until droplet formation

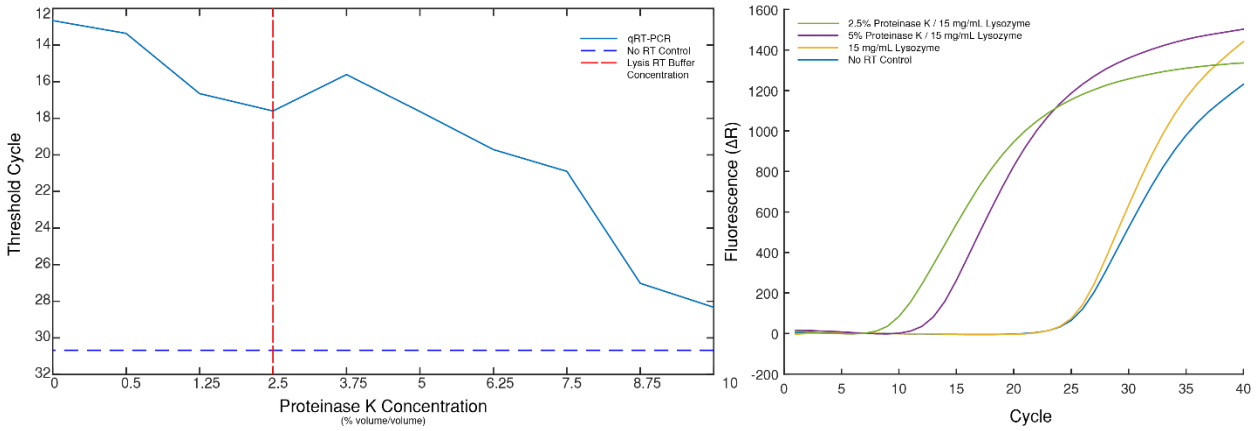



*Figure 93: Effects of Proteinase K and Lysozyme on RT efficiency. Left: qRT-PCRs were performed on RNA samples with various concentrations of Proteinase K added to the reaction. We see that RT efficiency begins to exponentially decay at Proteinase K concentraions above 3.75% (v/v). The red dashed line indicates the concentration of Proteinase K we use in our experiments. Right: qRT-PCRs were performed on RNA samples with the addition of Lysozyme, or lysozyme and proteinase K. The addition of Lysozyme alone to RT reactions results in no cDNA being generated, as the Lysozyme denatures and precipitates. The addition of proteinase K recovers cDNA synthesis as it is able to digest denatured Lysozyme.*

occurs, to prevent premature denaturation. Additionally, Proteinase K is added to our reaction solution to degrade lysozyme during the RT reaction, as lysozyme will also denature when the solution is heated to 50°C for reverse transcription. The proteinase K prevents lysozyme from precipitating and killing the RT reaction, but must be at a low enough concentration itself so as not to prevent the RT reaction directly. We stained cells with syto9 and then tested lysis of cells with only lysozyme in TE buffer as a positive control, our RT/Lysis Buffer as a test sample, and no lysozyme as a negative control to ensure that our RT/Lysis buffer was lysing bacterial cells (Figure 94). We found a similar pattern of cellular debris aggregating only in our test sample and positive control. We also tested our lysis buffer using *E. coli* cells expressing mCherry on a plasmid as we had previously found that mCherry was not aggregating, and we found that in our Lysis/RT solution mCherry had visibly diffused throughout the droplet, while in our negative control it had not (Figure 95). However we still see whole cells in our Lysis/RT solution in this test, which indicates that we do not have complete lysis. However, due to the throughput of this technique, any non lysed cells will not appear in the down-stream sequencing data and will simply act as empty droplets, and as the sequencing depth is currently limiting, a lysis rate of ~60-80% is acceptable. Determining a precise lysis rate is difficult as the size of the droplets is much larger than the size of the bacterial cells, and thus not all cells can be in focus at a single time.

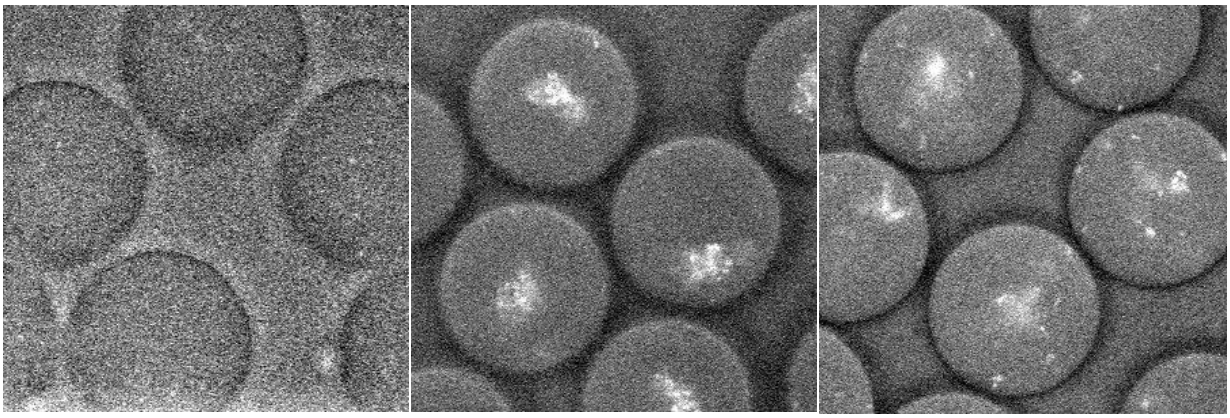


*Figure 94: Testing Cell Lysis effectiveness with Syto9 Reporter. MG1655 cells were stained with Syto9 and encapsulated into droplets containing and no lysozyme (left), lysozyme in TE buffer (centre), or RT/Lysis solution used in our Drop-Seq protocol (right). We find that lysed cells aggregate in the positive lysozyme TE buffer test and in our Drop-Seq protocol, but not in the negative no lysozyme control.*

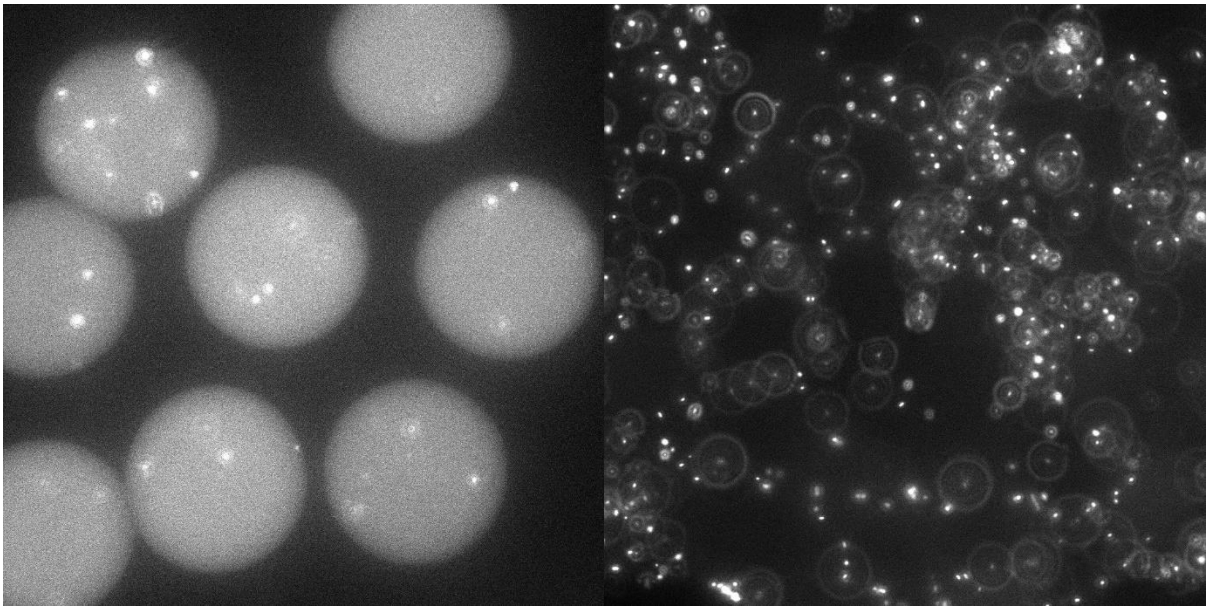


*Figure 95: Testing Cell Lysis efficiency with mCherry reporter. E. coli cells expressing mCherry on a plasmid were encapsulated with Drop-Seq protocol (left) or with the Drop-Seq Protocol without Lysozyme (right). We used a very high lambda to ensure that cells would be visible. When lysozyme is present, lysed cells release mCherry into the droplet and the droplet is clearly visible. It is noted that not all cells are currently lysed.*

***Bacterial cDNA is purified from unused primers and genomic DNA with biotinylated dCTP***

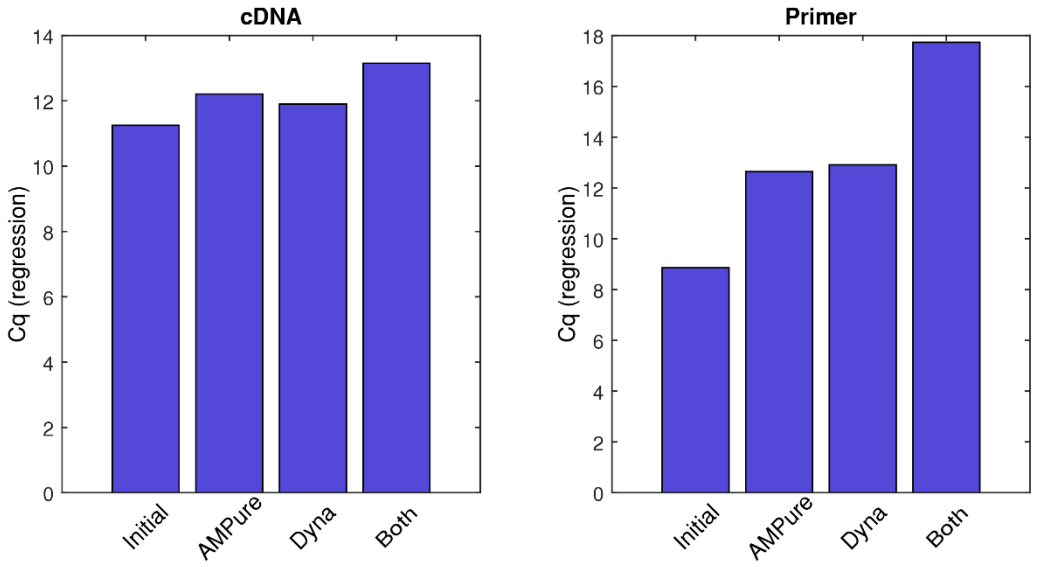



*Figure 96: Removal of excess RT primers with AMPure and Dynabeads. qRT-PCR was performed on RNA-spikes, with a biotin-14-dCTP integrated into the cDNA during synthesis. The RT-PCR was spiked with a DNA oligo that was the same size as primers used for droplet-based RT-PCR, and at approximately the same concentration. The cDNA was then purified with their size exclusion, using AMPure beads, with streptavidin coated magnetic Dynabeads, or both. The concentration of the cDNA and the concentration of the spiked primer-like oligo were quantified by qPCR.*

Because we cannot amplify our cDNA across exon junctions, or purify our RNA in droplets prior to RT, we need to separate our genomic DNA from our cDNA after RT in droplets is complete to prevent contamination. Therefore we perform our RT reaction using a biotinylated dCTP. This results in all of our cDNA generated to become biotinylated, but not our RT primers. We found that for effective incorporation of biotinylated dCTP we needed to use a dNTP mix with a reduced concentration of dCTP (0.2 mM instead of 0.5 mM) and biotinylated dCTP at a concentration of 0.1 mM. Additionally, we found the use of an 11 carbon linker had much higher incorporation rates than a 14 carbon linker. cDNA purification is performed by breaking RT emulsions with perfluoro-octanol. The aqueous phase is removed and ran on a PCR clean up column to remove non-incorporated nucleotides. 25uL of Streptavidin coated magnetic beads are added to the RT solution and incubated at room temperature for 3 hours. The magnetic beads are then washed and the cDNA is freed from the beads with 0.1% SDS solution at 99°C for 15 minutes. We found that we were able to recover approximately all of the cDNA in this technique, though we also had significant non-specific binding to the magnetic beads (Figure 96). Non-specific binding can be improved with addition of BSA (5 ug/mL) to binding buffer and addition of Urea (2M) to the washing buffer.

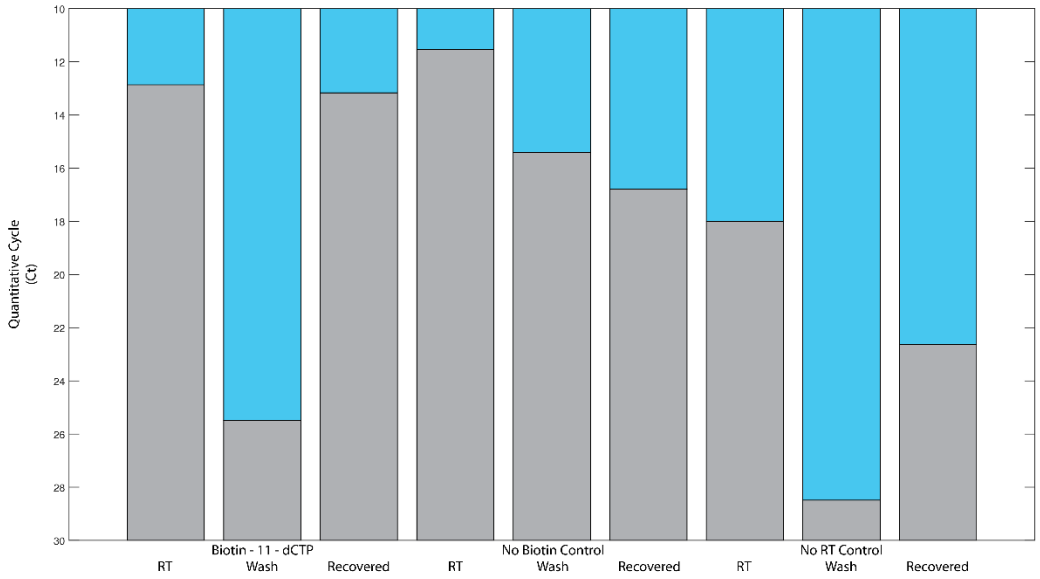



*Figure 97: Recovery of biotinylated cDNA with hydrogel beads. qRT-PCR was performed with biotinylated dCTP (Biotin-11-dCTP), without biotin (No Biotin Control), and without reverse transcriptase (No RT Control). After RT-PCR, cDNA was split into 2 equal volumes, one was used directly for qPCR (RT) and one was purified using Streptavidin coated magnetic beads (Recovered). Washes from magnetic beads were saved and tested for qPCR to determine if any cDNA was not captured (Wash). Biotinylated dCTP was successfully recovered using Streptavidin coated magnetic beads as indicated by the RT and Wash qPCRs having the same quantitative cycle. The cDNA recovered in the no biotin control indicates the level of non-specific binding.*

## 5.3 Single Cell RNA sequencing of MG1655

***Only one gene makes up the majority of unique UMIs for each cell***

To test the effectiveness of RNA capture in droplets using our protocol, we tested it on MG1655 *E. coli* cells. We designed a panel of genes that would give us an idea of which genes we could detect by taking 12 bulk RNA-seq experiments, normalizing them to have the same median count, and then taking the mean read count for each gene from all 12 experiments. Figure 100 shows that most genes had a mean read count of approximately 300. We then took the gene with the smallest standard deviation from each bin of our histogram as a reporter gene for that expression level. We also took all 5 genes in the second largest bin to ensure that we would be able to detect something if the system was working. We then designed gene specific primers for all 20 of these genes to determine how sensitive the protocol was (Figure 101). Our first sequencing experiment focused on only the 5 highest bins of expression levels to limit the possibility of primer dimers occurring in downstream PCR reactions. We grew *E. coli* strain MG1655 to an OD of 0.05 in M9 Media supplemented with 0.4% glucose, and performed our modified inDrop single cell RNA sequencing. The resulting barcoded cDNA was sequenced on a HiSeq with ~200 million clusters.

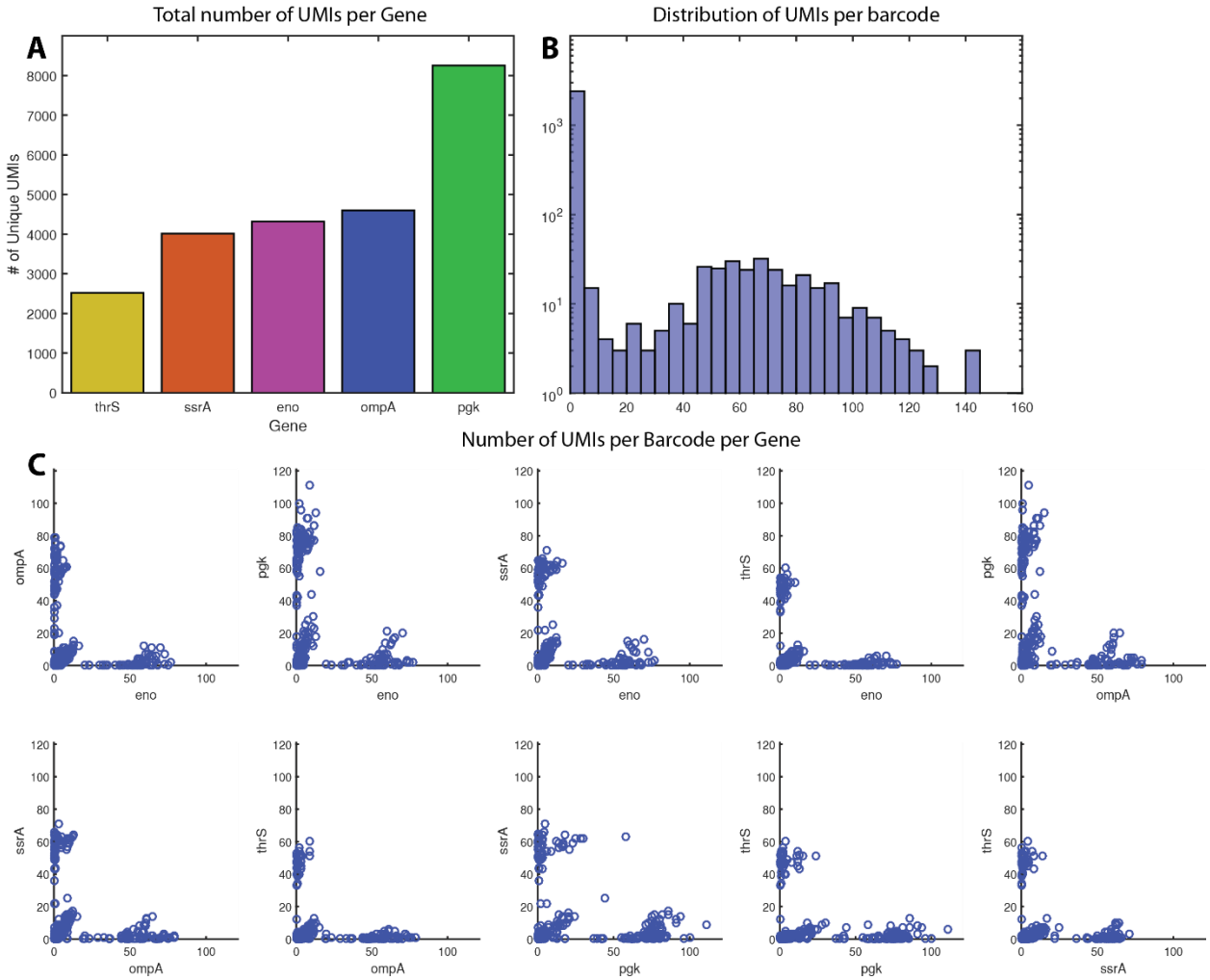



*Figure 98: Sequencing Results after relaxed filtering. (A) The total number of UMI detected for each gene. (B) The distribution of the number of UMIs detected for each barcode. (C) The number of UMIs for each pair of genes for each barcode.*

When we plotted a histogram of the reads per UMI of our data (weighted by the number of reads) we obtained the expected distribution (Figure 102A). However the scale was different than what we had expected from mammalian cell analysis by a few orders of magnitude. We therefore used a log scale to see if there were other populations of barcodes hidden in the histogram by the weighting technique used, and found that we had 3 populations of UMIs (Figure 102B). The left-most tail, with a few reads per UMI is likely PCR and sequencing errors creating new UMIs. However we have not identified the middle population. We filtered out the left most reads and plotted the combined unique UMIs for each gene in all barcodes (Figure 98A). The panel of genes followed our bulk expectation with the exception that the position for ssrA and pgk were switched. We plotted a histogram for the number of UMIs per barcode and found that we still had a left sided tail with many barcodes only having a few UMIs, but then a wide normal distribution centered at ~70 UMI per barcode (Figure 98B). Finally, we plotted the number of unique UMIs of each gene (in pairs of genes) for each barcode (Figure 98C). Rather than having genes being correlated between cells, which would indicate primarily variation in RNA capture, we found that cells where a reasonable number of RNA were detected (>20 UMI/Barcode), the RNA detected were almost all from the same gene. The scatter plots reveal 6 distinct populations in 5 dimensional space, one population corresponding to each gene, and a central population of barcodes with very few UMIs in total. To investigate if this was due to where we threshold our data, we made the same data plot with no threshold (Figure 103) and with a threshold just for the UMIs with the highest number of reads (Figure 104). In the raw data, the same pattern of gene expression was observed, but was exaggerated. Each individual cluster of barcodes was highly correlated in its own dimension. The pattern of the sum of all the genes detected was also similar to the initial, relaxed filtering. When we looked at the distribution of UMIs per barcode, again it had a left-handed tail in addition to a normal distribution, this time centered around 225 UMIs per barcode. Here though the left sided tail was enhanced compared to filtered data. Within the stringent filtering, we only observed 1 or 2 UMI per barcode. The sum of all UMIs per gene followed the same pattern, but the tiny number of UMI per barcode made the other analysis uninformative.

***PCR and sequencing errors are unlikely to explain highly patterned results***

This leaves us with a few possibilities. Firstly, the right hand population of a very high number of reads per UMI are the only true UMIs, and the other UMIs are all artifacts. It's possible that the large number of PCR cycles and that the PCR extends the sequence length during each step could contribute to strong PCR bias or artifacts. The long 5' extensions during the PCR do result in significantly different Tm (melting temperatures) for original and new templates. That is to say, it is much more likely to amplify a UMI which has already been amplified. Also the large number of cycles increases the chances of PCR errors being introduced. To check if the additional barcodes found with a more relaxed filter could be explained by PCR errors, we looked at the Levenshtein distance between barcodes found in the relaxed filtered (Figure 99A) and stringent filtered sets (Figure 99B). The levenshtein distance is the minimum number of changes required to change one sequence into another. The distribution of levenshtein distances shows peaks that are centered on multiples of 10, up to 40, with increasing proportions of the total barcodes. This is consistent with the data occurring from a random subpopulation of all possible barcodes. Each peak represents barcodes which are separated by a different number of barcodes, with the average distance between barcodes in each index approximately 10. Therefore the large population of barcodes which are ~40 levenshtein distance apart do not share any indexes, while those ~30 distance apart share at least one index. This largely excludes the possibility of PCR errors accounting for the additional barcodes. There is also the possibility of template switching during the PCR, in which recombination occurs

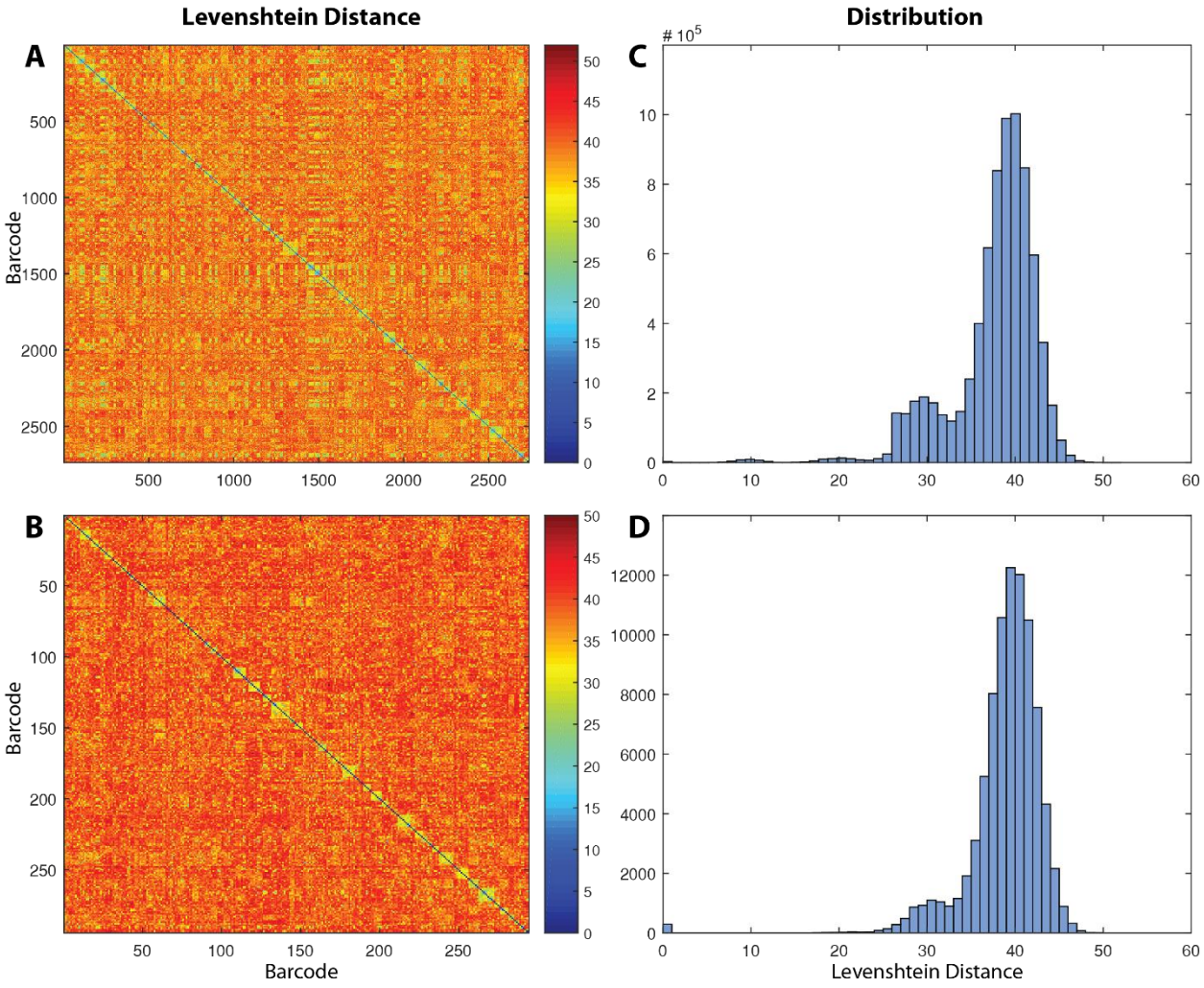



*Figure 99: Levenshtein distance between barcodes under different sequencing thresholds. The levenshtien distance between each pair of barcodes detected in scRNA sequencing is compared using a relaxed (A,C) and stringent (B,D) threshold for the minimum numbers of reads per UMI to be considered. If extra barcodes detected on relaxed threshold were due to sequencing errors, we expect a large enrichment of barcodes with very small Levenshtien distances (the number of changes required to make the two barcodes identical). In both cases, the majority of pairs have a large levenshtein distance, which is not explained by sequencing or PCR errors, but does not eliminate errors from template switching in PCR, or gene specific primer biases on hydrogel beads.*

between two unique sequences to give a new, apparently unique barcode. While this is usually considered a feature of reverse transcriptases but is known to happen in PCR as well [164]. We checked to see how many unique indexes there were in each population of the sequencing data, and we found sequences for all 4 indexes in the relaxed filtered set that were not found in the stringent filtered set. These barcodes could not arise from PCR recombination. Additionally, the short template size, low template number and long extension times used should severely limit the incidents of PCR recombination. Finally, carry over RT primers from the droplet RT-PCR could amplify in downstream PCRs. These can be limited through our stringent purification, or treatment with exonuclease I after RT-PCR to digest any remaining primers [159] however they can never be completely eliminated. These types of artifacts likely explain the left-sided tail on our distribution of the number of UMI per barcode, as we expect random primer carry over to have a few unique UMI and gene combinations, as there are hundreds of thousands to a million unique barcoded HgB per experiment, the probability that multiple carry over primers with the same barcode would amplify

in the PCR is low. In the relaxed filtered set it is important to note that individual barcodes (cells) do move off the axis, indicating that while they primarily have mRNA for a single read, they also have small amounts detectable for the other genes as well. This is not something we would expect from carry over primers with such a large diversity of barcodes. Finally, there is the possibility that our hydrogel beads are biased towards individual gene specific primers. If the majority of primers on a Hydrogel bead only bind to one gene, it would give the observed results. This is something that would not be observed with polyT primers for mammalian cells. We do premix our gene specific primers before they are added to the ligation mix and hydrogel beads, to reduce this possibility, though we cannot rule it out.

There is also a biological explanation for the observed single cell gene expression pattern. Bacterial mRNA only has a half-life of approximately 4 minutes, compared to upwards of 30 minutes with mammalian cells. This would result in a much stronger signal from transcriptional bursts. This is the observation that transcription occurs in short bursts of activity [165], and that the total relative number of transcripts in bulk sequencing assays is primarily driven by differences in frequencies of bursts rather than differences in maximum expression levels. Real-time monitoring of transcriptional bursts revealed they have duration of only 300-500 seconds in *E. coli*. This could be demonstrated with RNA-spikes as a control, as is done with single-cell qRT-PCR, however RNA-spikes are generally not compatible with droplet based single cell RNA sequencing, as the poisson distribution of 0.1 for cells means that the spikes will quickly consume sequencing reads, as they are present in all droplets while cells are present in only a few. We are therefore currently working on alternative controls, to identify the source of the signal, biological or technical, as well as optimize our barcoded hydrogel beads.

## 5.4 Single Cell analysis of Antibiotic Persister Cells

As the development of single-cell transcriptomics with bacterial cells is alone a complex and valuable technological development, we would like to demonstrate and validate this method independently from using it for high-throughput screening with CRISPR-Cas mutants. Therefore we will perform initial experiments for droplet based single bacterial RNA-seq to investigate variations in transcription patterns in *E. coli* sub-populations and how they contribute to Antibiotic Persistence. This is an ideal system for initial experiments as: (1) it does not require any cloning ; (2) it is clinically significant; (3) it is necessary to go to the single cell level to disentangle the phenotypic heterogeneity underlying persistence.

The growing prevalence of antibiotic resistance is a major problem confronting modern medicine. There are 400 000 infections caused by multidrug-resistant bacteria in Europe each year, and 25 000 of these cases are fatal [166]. Numerous national and international programs have been launched in response, often targeting resistance surveillance and the mechanisms of resistance propagation (e.g. EvoTAR: www.evotar.eu). This prevalence is due both to overuse of antibiotics and from exposure to lower sub-MIC concentrations in different, usually man-made, environments. Yet the acquisition of antibiotic resistance genes is not the only method that bacteria have to survive antibiotic treatments. Within a given bacterial population there exists a sub-population of dormant cells that are able to survive antibiotic treatments because the cellular processes that are typically corrupted by antibiotics are not active. Termed 'persisters', this sub-population is able to resume growth after the removal of antibiotics and are distinct from antibiotic resistant mutants as their progeny are just as sensitive to the antibiotic as the original population. As a result, antibiotic treatments must extend well past the alleviation of symptoms to prevent recurrent infections from these dormant cells. To complicate matters, low concentrations of antibiotics were shown to induce 'persister' cell formation, which then resist high, cytotoxic, concentrations of antibiotics [8]. Persistence, is a phenotypic switch that does not involve genotypic change, but has dramatic consequences in chronic infections [167] and favors the development of genetically acquired resistance itself. This mechanism is regulated by multi-factorial and incompletely understood regulatory pathways involving the stringent and SOS responses. As a result, antibiotic resistance and persistence are distinct but intertwined processes with significant clinical consequences.

The phenotypic heterogeneity that exists even within small cell populations prevents measurements based on the averages of bacterial populations to account for the small but sometimes critical changes rarely occurring in individual bacteria and leading to antibiotic persistence. Studying populations of bacteria at a single-cell level is thus essential to fully understand the physiological basis of antibiotic persistence. Indeed, individual cells can differ dramatically in size, protein levels, and expressed RNA transcripts. The crucial role of RNAs relies both in the fact that they reflect genome expression and in the evidence showing that they are also likely to impact genome function as well. Therefore, to answer previously irresolvable questions biologists would ideally like to map changes in RNA levels from single bacteria within a population. This would allow interactions between multiple redundant proposed pathways to be deciphered and to quantify to what extent different pathways contribute to the overall population structure. Our approach for performing RNA-sequencing on single bacterial cells in microfluidic drops has immediate applications to this problem. We thus aim to use this technique for bacterial transcriptomics analysis, first using *E.coli* and then the clinically relevant pathogen, *Vibrio cholerae*, selected following the use of commonly used antibiotic treatments. This powerful technology will lead,

for the first time, to high throughput investigation of bacterial persistence analysis that heretofore have been impossible to perform.

Over sixty years after its discovery [168], persistence has been thoroughly investigated to uncover the general mechanisms and associated genes, however almost all studies have been performed on bulk bacterial cultures [169] [170]. Variation of the response extent at single cell level is almost unknown and has made it difficult to disentangle the multiple pathways that can lead to persister phenotypes. Given the random, transient and heterogeneous character of persistence, it is absolutely necessary to qualify the underlying gene expression pattern at the single cell level [171]. Bacterial pathogens such as *V. cholerae*, or *Klebsiella pneumoniae*, when incubated in presence of all classes of antibiotics at sub-inhibitory concentrations varying from 1 to 10% of the MIC, trigger an SOS response and enter a mutagenic state [172] [173]. Preliminary RNA-Seq data suggests that these concentrations also trigger the stringent response in *V. cholerae*, which can explain the formation of persisters observed after sub-MIC aminoglycoside or ciprofloxacine treatments.

Next generation sequencing (NGS) generates massive amounts of data and has the potential to provide large amounts of gene expression analysis at the single cell level. Currently hundreds of strains can be analyzed in a single run and transcriptomics [RNA-Seq], which produces millions of reads that are used to quantify cellular gene expression. In parallel, single-cell equipment (e.g. Fluidigm C1) has been recently developed and commercialized to facilitate access to analyses at the level of single cells. However these machines are primarily well suited for eukaryotic cells and the application to small and potentially motile bacteria remains problematic. Furthermore, costs and total throughput remain outstanding issues. Droplet-based microfluidics coupled to NGS has been recently used to overcome these limitations, allowing single cell transcriptomics of embryonic stem cells [17] and genome wide expression profiling of mammalian cells [18], of several thousands of cells in a single run. The combination of droplet microfluidics and Next Generation Sequencing still needs to be adapted and applied to prokaryotes. This will offer a new powerful tool to understand antibiotic persistence phenotypes and open other new avenues of research.

The existing knowledge regarding antibiotic persistence makes it an ideal candidate for initial experiments. Many molecular pathways have been implicated in *E. coli* and genetic targets to reduce persistence have already been identified. Activation of Toxin Antitoxin modules leads to growth arrest by interfering with translation. However there is current disagreement as to the upstream pathways which activate these Toxin-Antitoxin modules, or if upstream pathways are required at all. Current methodologies are largely limited to bulk measurements, and it remains unknown how much each potential pathway contributes to the overall persister phenotype, and how these pathways interact. We have identified 23 key genes to screen for persister phenotypes, these include the aforementioned toxin genes, genes involved in the stringent response and implicated in the degradation of the antitoxins, genes involved in both the early and late stage SOS response, and genes recently identified in the flagella and serine biosynthesis pathways indicated to reduce persistence phenotype when deleted.

The fraction of *E. coli* cells that exist in a persister phenotype is correlated with the growth phase of the culture, therefore we plan to analyze single cells taken from both exponential growth phase (low persisters) and stationary phase (high persisters). Additionally, as persisters are a rare phenotype, we plan to enrich a culture for persister cells. Persister cells can be enriched in a population by treating with ampicillin as demonstrated by Lewis previously for transcriptional analysis of *E. coli* persister cells with

microarrays. Due to the lysis of non-persister cells by ampicillin, *E. coli* cultures treated with ampicillin can be centrifuged to recovered unlysed persister cells [174]. An alternative approach to enrichment of persister cells is to pre-screen them with FACS. As dormant cells are expected to have a low rate of protein synthesis, a strain that carries an unstable variant of GFP under the control of the ribosomal rrnB P1 promoter can be used to sort cells with a low fluorescence and thus a low rate of protein synthesis. After sorting Lewis reported a 5 fold increase in the percentage of cells that were persistent to ofloxacin [175].

In parallel, all cultures screened with microfluidics will also be subjected to CFU (colony forming unit) assays with a set of commonly used antibiotics (ampicillin, ciprofloxacin, and gentamycin) to classically determine the fraction of persister cells in each sample. Multiple antibiotics will be utilized as it has previously been demonstrated that persisters for one class of antibiotic are not necessarily persistent for all antibiotics, and to determine if persisters for specific classes of antibiotics correspond to specific molecular pathways. The well characterized nature of the molecular pathways in *E. coli* allows us the opportunity to explore how pathways implicated in persistence in *Vibrio cholerae* interact with other potential persister pathways. Because exposure of sub-MIC levels of antibiotics can induce persister formation through activation of the SOS response and the Toxin higB, we will also screen *E. coli* cultures that have been stressed with sub-MIC levels of ampicillin, ciprofloxacin and gentamycin to observe how the other implicated persistence pathways respond to antibiotic treatments, as initial results from our Partners at Institute Pasteur shows that this is of critical importance in *Vibrio cholerae*.

## 5.5 Methodology

*Photolithography and Soft Lithography for creation of microfluidic devices*

Photolithography was used to create silicon wafers with the desired microfluidic design on the surface. Wafers were heated at 200°C for 2 minutes to remove any vapor from the surface before adding a layer of Resin SU-8 at the desired thickness using a spin coater. For the Bacterial InDrop single cell RNAseq device, the channel height was approximately 40 µM. For this thickness, the spin coater was set to 2000 rpm for 30 seconds, soft baked for 3 minutes at 65°C and then for 5 minutes at 95°C. The wafer was then exposed to UV light (385nm negative photoresist) for 50 seconds using a Manual mask aligner MJB4, and baked again for 2 minutes at 65°C and for 5 minutes at 95°C. The wafer was then developed for 3 minutes.

Softlithography was performed by mixing Sylgard 184 polydimethylsiloxane (PDMS) base with the curing agent in a ratio of 9:1, consisting of 45g of base and 5g of curing agent for each microfluidic device. The mixture was then poured over the silicon wafer master mold, and placed in a vacuum chamber for 30 minutes to remove dissolved gases from the PDMS mixture. The PDMS was then baked at 70°C for 2 hours. After hardening, the PDMS was removed from the silicon wafer and trimmed to the appropriate size with a scalpel, and holes for the inlets and outlets were punched using a WellTech Rapid-Core 0.75 biopsy punch. The PDMS and a Corning plain glass micro slide were cleaned using scotch tape to remove any dust. The clean glass slide and PDMS were then placed into a Diener Zepto Plasma Cleaner with the microfluidic channels facing upward on the PDMS. The slide and PMDS were then treated with oxygen plasma at ~35 W for ~30 seconds. The PMDS and slide were removed from the plasma cleaner and the two exposed surfaces joined together to covalently bond the PDMS to the glass slide. The microfluidic device was allowed to rest for 15 minutes for the bonding to occur, afterwards the device was silanated by flowing 1% 1H,1H,2H,2H-Perfluorooctyltriethoxysilane in HFE-7500 oil through the microfluidic device. Excess oil and silane solution was removed with nitrogen gas, and then rinsed with HFE-7500 oil, which again was removed with nitrogen gas.

*Preparation of Hydrogel Beads (*LBC PEG-DA Bead protocol)

Hydrogel beads are created from a polyethylene glycol diacrylate solution containing an oligonucleotide with an acrydite 5' modification. This modification allows for the DNA to be covalently incorporated into the hydrogel bead. A solution containing 100 µL of 400 µM Acrydite dsRandomA, 93 µL 10% (w/w) PEG-DA-6000, 769 µL Tris 0.1 mM pH 8, 8.9 µL 1% (w/w) PEG-DA-700, 20 µL 2 µM FITC, 9 µL 1% (v/v) photo-initiator (2-hdyroxy-2-methylpropiophenone) was created in a microcentrifuge tube protected from light with an aluminum foil covering. Mixture was vortexed and centrifuged for 5 minutes at 11,000 RCF. Solution was loaded into a 1 mL glass syringe covered with black tape to protect the solution from light. A 5 mL glass syringe was loaded with HFE 7500 fluorinated oil containing 2% (w/v) Krytox surfactant. Syringes were loaded into separate Harvard syringe pumps and connected to a microfluidic droplet maker with tubing containing an interior diameter of 0.34 µM. Tubing was covered in black tape for the hydrogel solution to protect from light. Initial flow rates were 150 µL/h for the hydrogel solution, and 500 µL/h for the oil solution. Microfluidic device was placed within a microfluidic station to record the droplet volume. Hydrogel and Oil solution flow rates were adjusted until droplet volume was approximately 9 pL. Droplets were collected in tubing that past repeatedly (3x) under a UV lamp (360 mW, 365 nm) to initiate polymerization before being collected in a 5 mL eppendorf tube. Critically, the microfluidic chip must be protected from UV light during droplet production. Approximately

70 million hydrogel beads were generated for each 1 mL of hydrogel solution. After Bead production, hydrogel beads were washed by brief centrifugation at 800 RCF and removal of oil solution. Beads were washed with fresh HFE 7500 oil, briefly centrifuged at 800 RCF and the oil was removed again. Hexane was added to hydrogel beads up to a total volume of 5 mL. Hydrogels were mixed in hexane by pipetting until pellet was completely broken. Hydrogel beads were centrifuged for 10 seconds at 800 RCF to separate hexane from hydrogel beads and the hexane was removed. Hydrogel beads were then rinsed by adding washing buffer consisting of 0.1 M Tris-HCl and 0.1% Tween20 and mixing by pipetting. Hydrogel beads were then centrifuged for 3 minutes at 2500 RCF and the washing buffer was removed. The rinse with washing buffer was repeated 3 times. The beads were filtered by mixing approximately 20 million hydrogel beads (800 µL) with 15 mL of Binding and Washing buffer (20 mM Tris-HCl pH 7.5, 50 mM NaCl, 0.1% Tween20) and filtering through a 20 µm Millipore filter. An additional 15 mL of Binding and Washing buffer was added to wash filters. Filtered beads were centrifuged for 3 minutes at 2500 RCF, the supernatant was removed, and beads were re-pooled in a 5 mL eppendorf tube. Binding and Washing buffer was added to beads in a volume up to 5 mL. Beads were stored at 4°C.

*Generation of DNA Barcode Library*

Critically the split-pool process used to add DNA barcodes was done within a laminar hood to minimize contamination and low retention filtered tips were used to avoid loss of hydrogel beads. Approximately 10 million of filtered hydrogel beads (250 µL) were washed 3 times with 4 mL of Binding and Washing buffer (20 mM Tris-HCl pH 7.5, 50 mM NaCl, 0.1% Tween20). Each washing step consisted of adding 4 mL of Binding and Washing buffer, mixing well with a pipette, centrifuging for 2 minutes at 3000 RCF, and finally discarding the supernatant. The volume of hydrogel beads was then marked on the tube to make it easier to find the height of the hydrogel beads in downstream steps. The first adaptor was ligated to the hydrogel beads by making a solution of 1000 µL 2x T7 DNA Ligase Buffer, 20 µL of T7 DNA Ligase, 160 µL of 50 µM ds 5Pi0t-Photo-Rd1-iA (oMD381, oMD418, oMD419), and 570 µL Nuclease free water. The ds 5Pi0t-Photo-Rd1-iA contains a photo cleavable linker which allows it to detach from the Hydrogel Beads when exposed to UV light, but as a result, requires all further steps to be done in red light, and samples to be covered whenever possible to prevent premature cleavage of the linker. The Ligation solution is added to the washed Hydrogel Beads, mixed, and incubated for 30 minutes at room temperature in a tube rotator. After incubation, Hydrogel Beads were centrifuged for 2 minutes at 3000 RCF and the solution was removed from the beads. The beads were then washed as above, with the exception of the final wash, which contained only 750 µL of Binding and Washing buffer instead of 4 mL, and a 4 µL aliquot was taken before centrifugation during this final washing step and saved for later quality control. The first barcode oligonucleotide was ligated by making a ligation solution containing 1000 µL of 2x T7 DNA Ligase Buffer, 20 µL T7 DNA Ligase, and 340 µL of Nuclease free water and adding it to the hydrogel beads. The hydrogel beads were then mixed, and aliquoted into 12 wells of a PCR strip, of approximately 132 µL of Hydrogel Bead and Ligation solution each. The beads are then aliquoted into a 96 deep well DNA LoBind plate which is prefilled with 4 µL of 20 µM ds index A barcode oligonucleotides. A multichannel pipette was used to add 16 µL of Hydrogel Bead and Ligation solution to each well. The microtiter plate was covered and incubated for 15 minutes at 25°C while shaking at 600 rpm. The plate was removed from agitation and incubated for a further 10 minutes at room temperature before heat inactivation at 65°C for 10 minutes. After being allowed to return to room temperature, the plate was placed on ice and 200 µL of cold Binding and Washing buffer was added to each well. The solution from each well was transferred into 4 cold 5 mL DNA LoBind Eppendorf tubes and centrifuged at 4°C for 2

minutes at 3000 RCF. The supernatant was discarded and a fresh 2 mL of cold Binding and Washing buffer was added to each tube. The solution from two of the tubes was then pooled with one of the remaining two tubes. The two tubes containing solution were then centrifuged at 4°C for 2 minutes at 3000 RCF, and the supernatant was removed. For each tube, 375 µL of cold Binding and Washing buffer was added and then the two solutions were pooled together. A ligation control was taken by removing 4 µL of solution and saving it for quality control. This process of split-ligation-pooling was repeated for indexes B, C, and D. Gene specific primers were ligated to the hydrogel beads by making a solution of 1000 µL 2x T7 DNA Ligase Buffer, 20 µL of T7 DNA Ligase, 160 µL of 50 µM double stranded gene specific oligonucleotides (oMD588, oMD590, oMD592, oMD594, oMD596), and 570 µL Nuclease free water. The gene specific oligonucleotides are first made double stranded by annealing with oMD428, pooling, and mixing well. The Ligation solution is added to the washed Hydrogel Beads, mixed, and incubated for 30 minutes at room temperature in a tube rotator. After incubation, Hydrogel Beads were centrifuged for 2 minutes at 3000 RCF and the solution was removed from the beads. The beads were then washed as with the first oligonucleotide linker, and again a 4 µL aliquot was taken before centrifugation during the final washing step and saved for later quality control. The 4 µL aliquots for quality control were then exposed to UV light for 3 minutes, and 6 µL of Nuclease free water and 2 µL of 6x Loading Dye was added to each aliquot. Each aliquot was then loaded onto a 2% agarose gel stained with 1X GelRed. The final ligation control was also diluted 1/10 and ran on a Agilent 4200 Tapestation HS DNA 1000 to quantify the proportion of full length barcoded RT primers.

*Single cell droplet-based RNA Sequencing of Bacteria*

*E. coli* MG1655 cells were grown overnight in M9 Media + 0.4% Glucose. The next morning, cells were diluted 1/5 in fresh M9 media + 0.4% Glucose and grown for 3 hours to an OD (600nm) of 0.05. Cells were pelleted by taking 1 mL of culture and centrifuging for 2 minutes at 11,000 RCF. Cells were then resuspended in 50 µL of RNA Later and incubated at room temperature for 5 minutes. Lysis RT solution was made by mixing 10 µL of Superscript III Reverse Transcriptase, 5 µL of RNaseOUT Recombinant Ribonuclease inhibitor, 10 µL Lysozyme in TE buffer at a concentration of 150 mg/mL, 5 µL of Polymyxin B solution, 5 µL of 0.5M Trizma (pH 7.5), and 15 µL $H_2O$. Cells were diluted by mixing 5 µL of Cells, 11.875 µL of 1M Tris-HCl (pH 8.4), and 8.125 µL $H_2O$. The Cell solution was then made by mixing 10 µL of diluted cells, 2.5 µL of SUPERase in RNase Inhibitor, 2.5 µL Qiagen Proteinase K, 10 µL dNTP mix (10 mM dATP, 10 mM dGTP, 10 mM dTTP, 2 mM dCTP), 5 µL Biotin-14-dCTP (1 mM), 7.5 µL 1M KCl, 1 µL 270 mM $MgCl_2$, 5 µL 0.1M DTT, and 6.5 µL $H_2O$. Hydrogel bead loading solution was made by mixing 500 µL of 1M Tris-HCl (pH 8.4), 750 µL 1M KCl, 28 µL 1M $MgCl_2$ 7.71265 mg DTT, 500 µL 10% IPEGAL, and 8.222 mL $H_2O$. This solution was used to wash 50 µL of Barcoded hydrogel beads 3 times by centrifuging hydrogel beads for 1 minute at 4000 RCF, removing supernatant, adding 200 µL of Hydrogel bead loading solution, mixing well, and repeating. Three omnifit syringes were filled with HFE-7500 oil. Approximately 1m of tubing was attached to each syringe and filled with HFE-7500 oil by pushing on the syringes. 50uL of each solution above (Lysis RT, Cell and Hydrogel Beads) were loaded into the tubing of a syringe. Flow rates used were 200 uL/h for each solution, and 300 uL/h for HFE-7500 with 2% Surfactant. This resulted in drops with a volume of ~100 pL. Droplets were collected until solutions were consumed. Droplets were then incubated at 50°C for 3 hours. The emulsion was then broken and treated with exonuclease I to remove excess primers. Dynabeads M280 were washed with binding washing buffer (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl) supplemented with BSA. Dynabeads were used to bind to cDNA with incorporated biotin-14-dCTP. DNA bound Dynabeads were washed with binding washing buffer supplemented with 2 M urea.

DNA was freed from Dynabeads by incubating the beads in 95% formamide + 10mM EDTA, pH 8.2 for 2 minutes at 90°C. Illumina adaptors were added to cDNA with primer extension. Sequencing was performed by Institute Pasteur.

## 5.6 Supplementary Figures

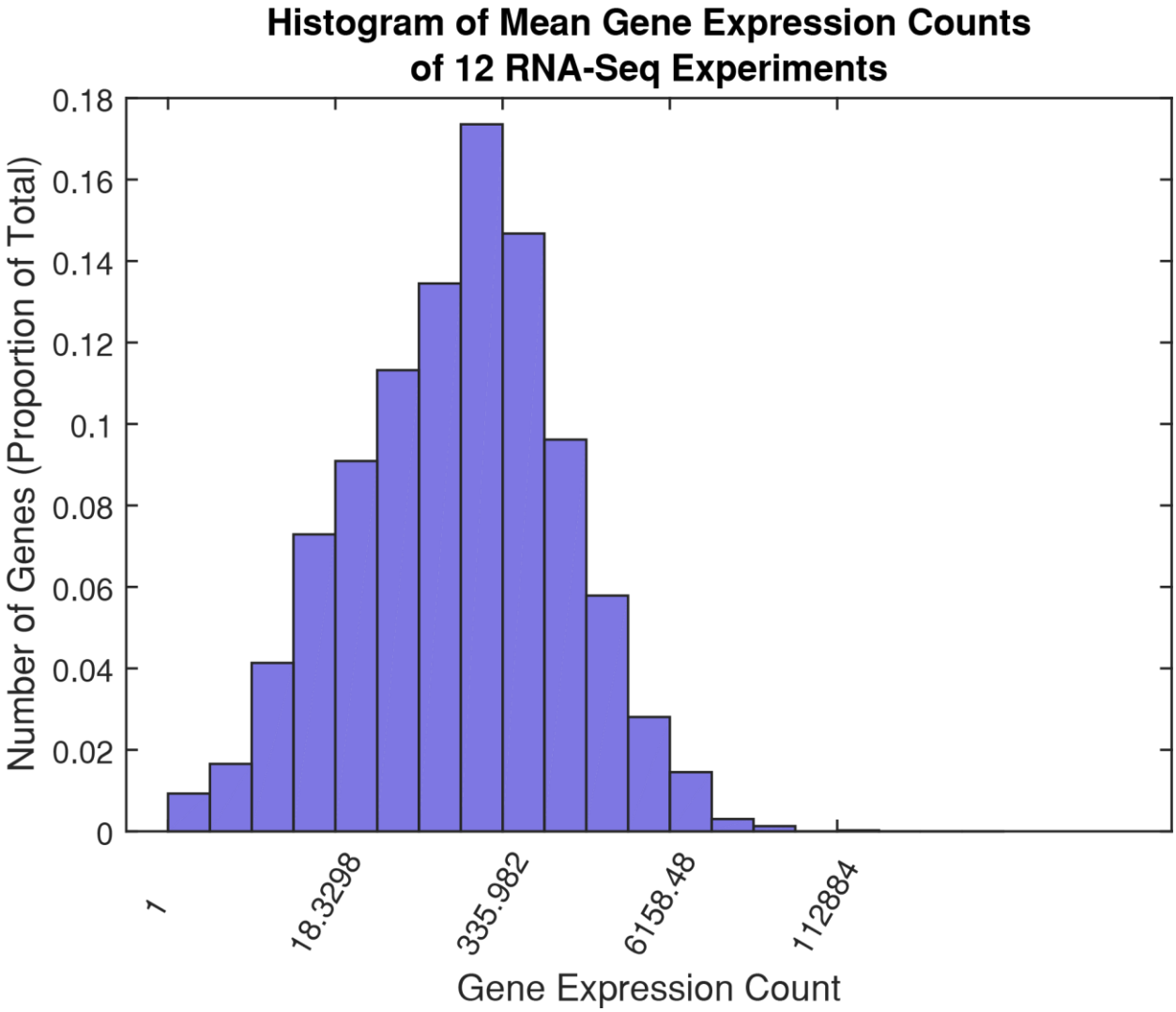



*Figure 100: Mean gene expression from 12 RNA-Seq experiments. RNA-Seq experiments were normalized to have the same median gene expression, and then the mean of each gene was taken. These were distributed into 20 bins determined by a logarithmic scale. The distribution of the mean expression for each gene indicates that most genes have between 150 to 350 reads per experiment.*

**Figure 31:**

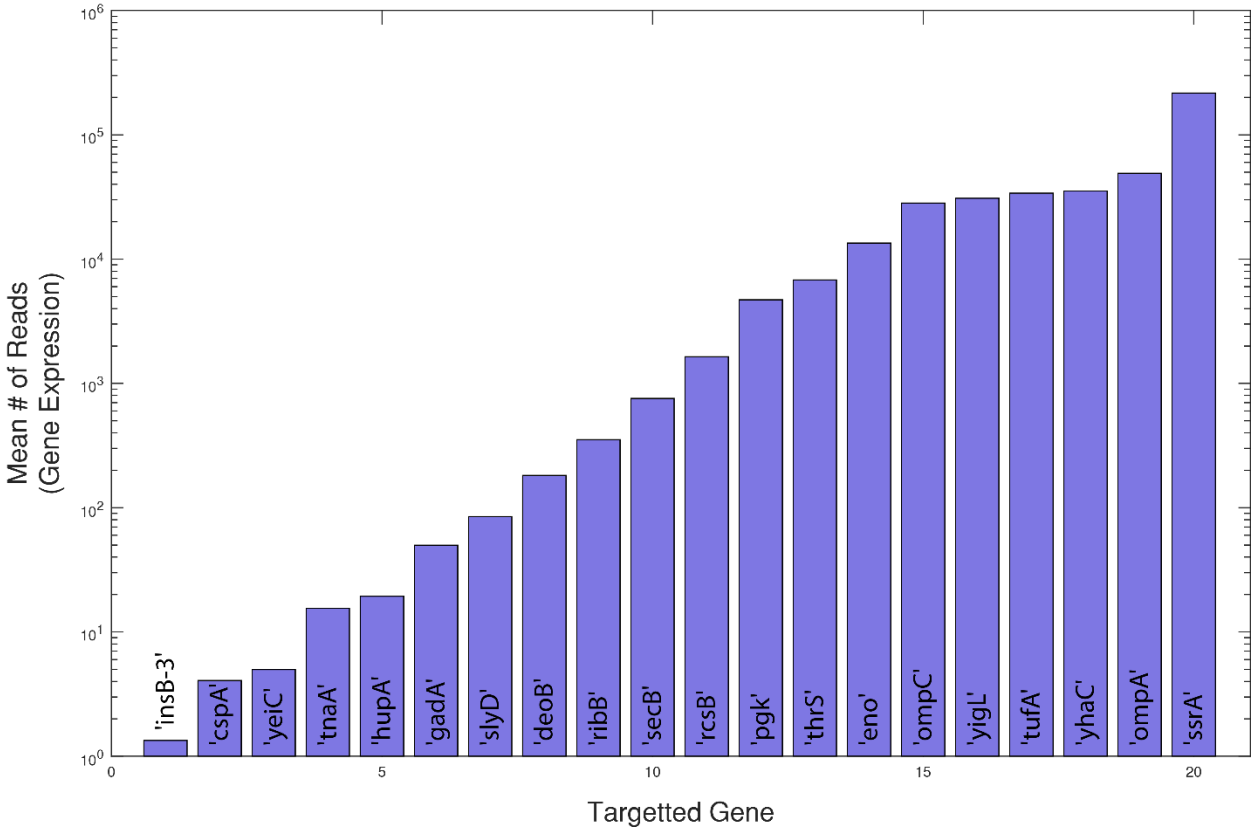



*Figure 101: Determining sensitivity of Drop-Seq using a panel of genes with a range of expression levels. Using the distribution of gene expression from Figure 30, one gene from each bin was chosen by the gene with the smallest standard deviation of expression within each bin. These 16 genes were chosen to be reporter genes to determine what the minimum level of expression we can expect to detect with single cell droplet based RNA-sequencing. We additionally included the remaining 4 genes from the second largest bin to ensure we had good detection of highly expressed genes for troubleshooting purposes. The mean number of reads expected from bulk RNA sequencing for each reporter is shown.*

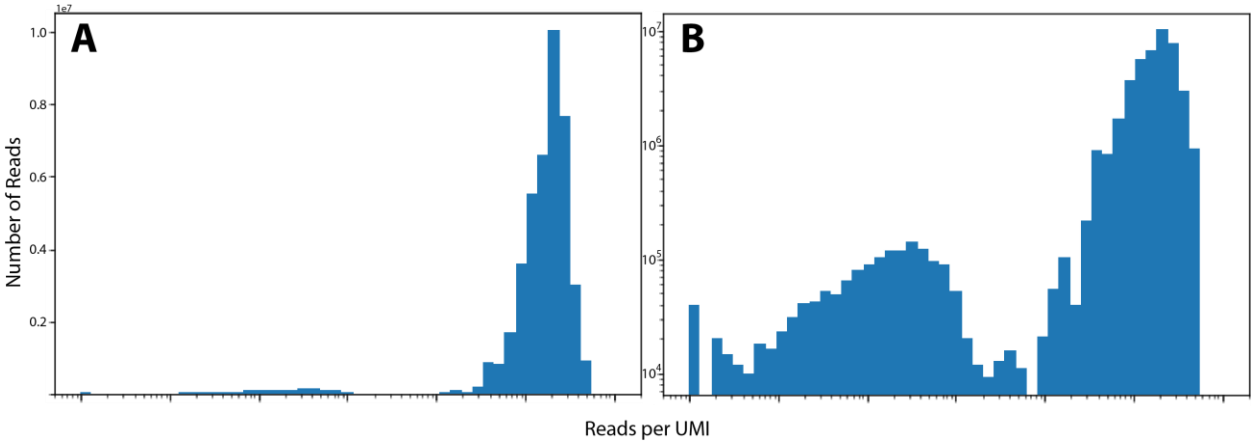



*Figure 102: Histogram of the number of reads per UMI. The histogram of the number of reads per UMI is weighted by the number of reads. In a linear scale (A) this gives the expected distribution based on previous experiments with mammalian inDrop techniques. In a log scale (B) there are three distributions detected.*

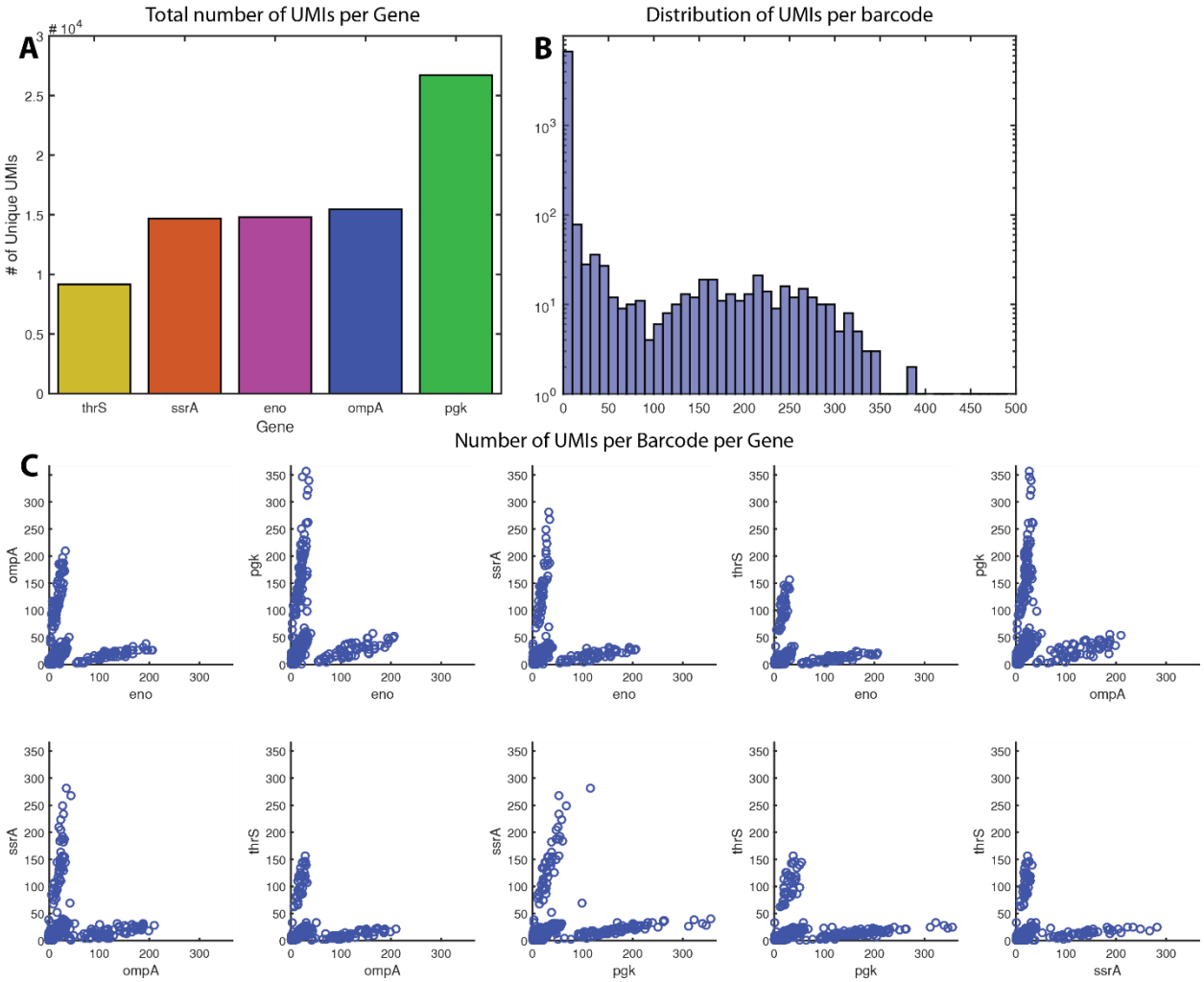



*Figure 103: Raw Sequencing Results. (A) The total number of UMI detected for each gene. (B) The distribution of the number of UMIs detected for each barcode. (C) The number of UMIs for each pair of genes for each barcode.*

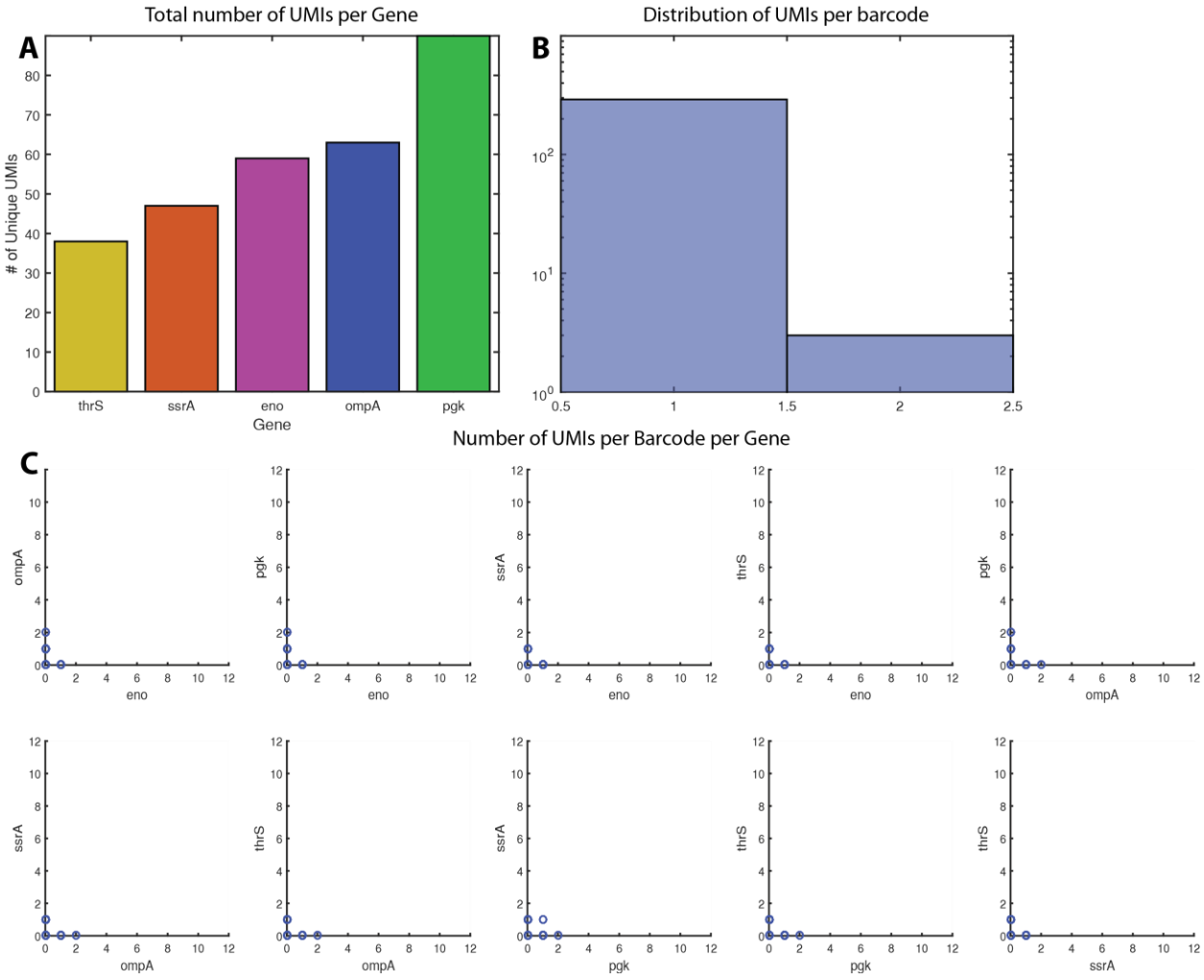



*Figure 104 Sequencing Results after stringent filtering. (A) The total number of UMI detected for each gene. (B) The distribution of the number of UMIs detected for each barcode. (C) The number of UMIs for each pair of genes for each barcode.*

# 6 Perspectives

Here we demonstrate a method for using CRISPR-Cas to generate high-dimensional perturbation libraries, and a potential method for screening them with droplet based microfluidics similar to recent methods for mammalian cells [19] [20]. There are two key differences between these approaches and the one detailed here. Firstly, the application for droplet-based single-cell RNA sequencing has not yet been demonstrated, and poses significant difficulties over mammalian cells. Secondly, CRISPR perturbations with pCKDL are uniform, cover all possible combinations of perturbations, and contain no redundancies in the perturbation library. This has significant advantages over relying on rare and random multiple infections with lentiviruses. Primarily, with a low multiplicity of infection used in Perturb-Seq, the proportion of double and triple perturbations is low compared to the number of single perturbations, even though this represents a much larger proportion of the total possible perturbations. Additionally, increasing the multiplicity of infection would result in increased instances of recombination between perturbations and barcodes, thus confounding the results. Finally, there are no perturbation combinations above triples, limiting the ability to detect higher order epistasis.

This technique was applied to the global transcription regulators of *E. coli*. We found significant higher order epistasis between these regulators. While much of the variance in gene expression can be attributed to the independent action of the global regulators, the dimensionality needed to explain most of the variance is very high. High order epistasis has been found previously with genotype to phenotype maps, and consisted of 2.2% to 31% of the variance in the data. We have additionally found that sign epistasis, which may restrict evolutionary trajectories, is dependent on both the selection pressure and the environment. By varying either of these, it is possible to project an otherwise rugged adaptive landscape into more dimensions, which can allow the organism to evolve around or out of otherwise prohibited fitness valleys. Taken together, the additionally dimensionality in the genetic response provided by higher order epistasis may provide the same benefits.

This would help to explain changings in regulation in terms of evolution. Despite global regulators acting as highly connected hubs in the transcriptional network, they tend to be highly evolvable. Traditionally network hubs are thought to provide critical roles, and as such should be highly conserved. Yet there are many observations that changes in global regulation through global transcriptional regulators are some of the earliest and most common mutations when cells adapt to a new environment [31] [32] [33] [34] [35]. We are currently developing a framework which predicts that these global regulators and the network structure itself act as a tuning device for gene expression. Two analogies could be a fitting function, or an artificial neural network. In the analogy of a fitting function, each global regulator acts like a parameter in the function, and such a large number of parameters allows the function to fit nearly any desired trajectories. Similarly, in the analogy of an artificial neural network, each transcription factor acts as a neuron and each strongly connected component in the transcriptional regulatory network acts like a layer of the artificial neural network. This allows the network to take on nearly any desired function. As such, we can view our perturbation library as each having a distinct trajectory through the potential genotype space. Individual trajectories come closer or farther to some idealized phenotype as they pass through this space.

There is still much work that needs to be done to confirm such claims however. First and foremost, RNA-sequencing data must be reproduced and must be done for additional growth media. We expect this to be completed this fall, but it is necessary to both estimate the noise in our experiments and to identify potential outliers or artifacts. Secondly, Competition and Swimming fitness measurements from chapter two must be replicated in a less burdensome system, to see if it improves reproducibility. At the moment the data is too noisy and not reproducible enough to draw conclusions from. If the plasmid burden is indeed causing the reproducibility issues it opens a further avenue of interest, how does the cell decide when to switch off costly functions for the cell? This is unlikely to be something that we will address immediately but how and why certain biological replicates, and even specific cells within an experiment, switch between swimming phenotypes despite being clonal is potentially very rewarding.

In terms of the microfluidic aspects, initial single-cell RNA sequencing data raises the question of whether there is enough RNA in a single bacterial cell, and if it is stable enough, to extract any meaningful data for high-throughput screening or detection of distinct subpopulations. Single-cell RNA sequencing in mammalian cells is already notoriously noisy data. Mammalian cells contain approximately 360,000 mRNA per cell [176] while bacterial cells only contain ~3000 [177]. The low number of mRNAs and their short half-life may make it difficult to reconstruct an accurate representation of the state of the cell and will correlate very poorly with the protein content of the cell.

Finally, our molecular barcoding for antibiotic combinations is not limited to antibiotics. We could imagine screening combinations of environmental factors such as carbon sources and pH, or combinations of environments with genetic perturbations. This opens up further avenues for large environment, genotype, phenotype mapping, and the modular nature of the microfluidics means that improvements in microfluidic designs can quickly be incorporated into our system.

# 7 Bibliography


[1] F. S. Collins, "Medical and Societal Consequences of the Human Genome Project," *New England Journal of Medicine,* vol. 341, pp. 28-37, 1999.

[2] F. S. Collins, E. D. Green, A. E. Guttmacher and M. S. Guyer, "A vision for the future of genomics research," *Nature,* vol. 422, pp. 835-847, 2003.

[3] S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes and M. T. Weirauch, "The human transcription factors," *Cell,* vol. 172, no. 4, pp. 650-665, 2018.

[4] S. Yi, S. Lin, Y. Li, W. Zhao, G. B. Mills and N. Sahni, "Functional variomics and network perturbation: conneting genotype to phenotype in cancer," *Nature reviews genetics,* vol. 18, pp. 395-410, 2017.

[5] R. Gawne, K. Z. McKenna and H. F. Nijhout, "Unmodern Synthesis: Developmental Hierarchies and the Origin of Phenotypes," *BioEssays,* vol. 40, p. 1600265, 2018.

[6] Z. D. Blount, "The unexhausted potential of E. coli," *eLife,* vol. 4, p. e05826, 2015.

[7] N. K. Priest, J. K. Rudkin, E. J. Feil, J. M. H. Van Den Elsen, A. Cheung, S. J. Peacok, M. Laabei, D. A. Lucks, M. Recker and R. C. Massey, "From Genotype to phenotype: can systems biology be used to predict Straphylococcus aureus virulence?," *Nature Reviews Microbiology,* vol. 10, no. 11, p. 791, 2012.

[8] T. Dörr, K. Lewis and M. Vulić, "SOS response induces persistence to fluoroquinolones in Escherichia coli," *PLoS genetics,* vol. 5, no. 12, p. e1000760, 2009.

[9] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis and J. N. Hirschhorn, "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," *Nature Reviews Genetics,* vol. 9, no. 5, p. 356, 2008.

[10] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano and G. Stolovitzky, "Revealing strengths and weaknesses of methods for gene network inference," *Proceedings of the national academy of sciences,* 2010.

[11] T. F. C. Mackay, "Epistasis and quantitative traits: using model organisms to study gene-gene interactions," *Nature Reviews Genetics,* vol. 15, no. 1, p. 22, 2014.

[12] D. Risso, J. Ngai, T. P. Speed and S. Dudoit, "Normalization of RNA-seq data using factor analysis of control genes or samples," *Nature biotechnology,* vol. 32, pp. 896-902, 2014.

[13] K. Wood, S. Nishida, E. D. Sontag and P. Cluzel, "Mechanism-independent method for predicting response to multidrug combinations in bacteria," *Proceedings of the National Academy of Sciences,* vol. 109, no. 30, pp. 12254-12259, 2012.

[14] R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. S. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, U. Schopfer and G. S. Sittampalam, "Impact of high-throughput screening in biomedical research," *Nature reviews Drug discovery,* vol. 10, no. 3, p. 188, 2011.

[15] P. Kemmeren, K. Sameith, L. A. L. van de Pasch, J. J. Benschop, T. L. Lenstra, T. Margaritis, E. O'Duibhir, E. Apweiler, S. van Wageningen, C. W. Ko, S. van Heesch, M. M. Kashani, G. Ampatziadis-Michailidis, M. O. Brok, N. A. C. H. Brabers, A. J. Miles, D. Bouwmeester, S. R. van Hooff and F. C. P. Holstege, "Large-Scale Genetic Perturbations Reveal Regulatory Networks and an Abundance of Gene-Specific Repressors," *Cell,* vol. 157, no. 3, pp. 740-752, 2014.

[16] D. Bikard, W. Jiang, P. Samai, A. Hochschild, F. Zhang and L. A. Marraffini, "Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system," *Nucleic Acids Research,* vol. 41, no. 15, pp. 7429-7437, 2013.

[17] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz and M. W. Kirschner, "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells," *Cell,* vol. 161, no. 5, pp. 1187-1201, 2015.

[18] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev and S. A. McCarroll, "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets," *Cell,* vol. 161, no. 5, pp. 1202-1214, 2015.

[19] D. A. Jaitin, A. Weiner, I. Yofe, D. Lara-Astiaso, H. Keren-Shaul, E. David, T. M. Salame, A. Tanay, A. van Oudenaarden and I. Amit, "Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq," *Cell,* vol. 167, pp. 1883-1896, 2016.

[20] A. Dixit, O. Parnas, B. Li, J. Chen, C. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman and A. Regev, "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens," *Cell,* vol. 167, pp. 1853-1866, 2016.

[21] N. Geva-Zatorsky, E. Dekel, A. A. Cohen, T. Danon, L. Cohen and U. Alon, "Protein dynamics in drug combinations: a linear superposition of individual-drug responses," *Cell,* vol. 140, no. 5, pp. 643-651, 2010.

[22] O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh and U. Alon, "Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space," *Science,* p. 1217405, 2012.

[23] A. Barnard, A. Wolfe and S. Busby, "Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes," *Current opinion in microbiology,* vol. 7, no. 2, pp. 102-108, 2004.

[24] Y.-Y. Liu, J.-J. Slotine and A.-L. Barabási, "Controllability of complex networks," *Nature,* vol. 473, no. 7346, p. 167, 2011.

[25] P. Nghe, M. Kogenaru and S. J. Tans, "Sign epistasis caused by hierarchy within signalling cascades," *Nature communications,* vol. 9, no. 1, p. 1451, 2018.

[26] M. G. J. de Vos, F. J. Poelwijk, N. Battich, J. D. T. Ndika and S. J. Tans, "Environmental Dependence of Genetic Constraints," *PLoS Genetics,* vol. 9, no. 6, p. e1003580, 2013.

[27] D. F. Browning, D. C. Grainger and S. J. W. Busby, "Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression," *Current opinion in microbiology,* vol. 13, no. 6, pp. 773-780, 2010.

[28] S. Berthoumieux, H. De Jong, G. Baptist, C. Pinel, C. Ranquet, D. Ropers and J. Geiselmann, "Shared control of gene expression in bacteria by transcription factors and global physiology of the cell," *Molecular systems biology,* vol. 9, no. 1, p. 634, 2013.

[29] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins and T. S. Gardner, "Large-scale mapping and validation of Escherichia coli transciptional regulation from a compendium of expression profiles," *PLoS biology,* vol. 5, no. 1, p. e8, 2007.

[30] N. C. Wu, L. Dai, C. A. Olson, J. O. Lloyd-Smith and R. Sun, "Adaption in protein fitness landscapes is facilitated by indirect paths," *eLife,* vol. 5, p. e16965, 2016.

[31] S. F. Elena and R. E. Lenski, "Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation," *Nature Reviews Genetics,* vol. 4, pp. 457-469, 2003.

[32] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein and S. A. Teichmann, "Structure and evolution of transcriptional regulatory networks," *Current Opinion in Structural Biology,* vol. 14, no. 3, pp. 283-291, 2004.

[33] A. Crombach and P. Hogeweg, "Evolution of Evolvability in Gene Regulatory Networks," *PLoS Computational Biology,* vol. 4, no. 7, p. e1000112, 2008.

[34] Z. D. Blount, J. E. Barrick, C. J. Davidson and R. E. Lenski, "Genomic analysis of a key innovation in an experimental Escherichia coli population," *Nature,* vol. 489, pp. 513-518, 2012.

[35] D. I. Kisiela, M. Radey, S. Paul, S. Porter, K. Polukhina, V. Tchesnokova, S. Shevchenko, D. Chan, M. Aziz, T. J. Johnson, L. B. Price, J. R. Johnson and E. V. Sokurenko, "Inactivation of transcriptional regulators during within-household evolution of Escherichia coli," *Journal of Bacteriology,* pp. JB-00036, 2017.

[36] F. Jacob and J. Monod, "Genetic regulatory mechanisms in the synthesis of proteins," *Journal of Molecular Biology,* vol. 3, no. 3, pp. 318-356, 1961.

[37] A. Martı´nez-Antonio and J. Collado-Vides, "Identifying global regulators in transcriptional regulatory," *Current Opinion in Microbiology,* vol. 6, pp. 482-489, 2003.

[38] S. S. Shen-Orr, R. Milo, S. Mangan and U. Alon, "Network motifs in the transcriptional regulation network of Escherichia coli," *Nature genetics,* vol. 31, no. 1, p. 64, 2002.

[39] H.-W. Ma, J. Buer and A.-P. Zeng, "Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach," *BMC Bioinformatics,* vol. 5, no. 1, p. 199, 2004.

[40] X. Fang, A. Sastry, N. Mih, D. Kim, J. Tan, J. T. Yurkovich, C. J. Lloyd, Y. Gao, L. Yang and B. O. Palsson, "Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities," *Proceedings of the National Academy of Sciences,* vol. 114, no. 38, pp. 10286-10291, 2017.

[41] A. R. Abate, T. Hung, P. Mary, J. J. Agresti and D. A. Weitz, "High-throughput injection with microfluidics using picoinjectors," *Proceedings of the National Academy of Sciences,* 2010.

[42] I. Junier and O. Rivoire, "Conserved units of co-expression in bacterial genomes: an evolutionary insight into transcriptional regulation.," *PloS one,* vol. 11, no. 5, p. e0155740, 2016.

[43] S. Gottesman, "Bacterial regulation: global regulatory networks," *Annual review of genetics,* vol. 18, no. 1, pp. 415-441, 1984.

[44] I. M. Keseler, A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martínez, C. Fulcher, A. M. Huerta, A. Kothari, M. Krummenacker, M. Latendresse, L. Muñiz-Rascado, Q. Ong, S. Paley, I. Schröder, A. G. Shearer, P. Subhraveti, M. Travers, D. Weerasinghe, V. Weiss, J. Collado-Vides, R. P. Gunsalus, I. Paulsen and P. D. Karp, "EcoCyc: fusing model organism databases with systems biology," *Nucleic Acids Research,* vol. 41, no. D1, pp. D605-D612, 2012.

[45] S. Gama-Castro, H. Salgado, A. Santos-Zavaleta, D. Ledezma-Tejeida, L. Muñiz-Rascado, J. S. García-Sotelo, K. Alquicira-Hernández, I. Martínez-Flores, L. Pannier, J. A. Castro-Mondragón, A. Medina-Rivera, H. Solano-Lira, C. Bonavides-Martínez, E. Pérez-Rueda, S. Alquicira-Hernández, L. Porrón-Sotelo, A. López-Fuentes, A. Hernández-Koutoucheva, V. Del Moral-Chávez, F. Rinaldi and J. Collado-Vides, "RegulonDB version 9.0: high-level intergration of gene regulation, coexpression, motif clustering and beyond," *Nucleic acids research,* vol. 44, no. D1, pp. D133-D143, 2016.

[46] N. Le Novère, "Quantitative and logic modelling of molecular and gene networks," *Nature Reviews Genetics,* vol. 16, pp. 146-158, 2015.

[47] H. Bolouri, "Modeling genomic regulatory networks with big data," *Trends in Genetics,* vol. 30, no. 5, pp. 182-191, 2014.

[48] R. E. Kalman, "Mathematical Description of Linear Dynamical Systems," *Journal of the Society for Industrial and Applied Mathematics Series A Control,* vol. 1, no. 2, pp. 152-192, 1963.

[49] D. G. Luenberger, Introduction to dynamic systems: theory, models, and applications, New York: Wiley, 1979.

[50] J.-J. E. Slotine and W. Li, Applied nonlinear control, Englewood Cliffs NJ: Prentice hall, 1991.

[51] L. Marucci, D. A. W. Barton, I. Cantone, M. A. Ricci, M. P. Cosma, S. Santini, D. di Bernardo and M. di Bernardo, "How to turn a genetic circuit into a synthetic tunable oscillator, or a bistable switch," *PloS one,* vol. 4, no. 12, p. e8083, 2009.

[52] N. J. Cowan, E. J. Chastain, D. A. Vilhena, J. S. Freudenberg and C. T. Bergstrom, "Nodal dynamics, not degree distributions, determine the structural controllability of complex networks," *PloS one,* vol. 7, no. 6, p. e38398, 2012.

[53] O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh and U. Alon, "Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space," *Science,* p. 1217405, 2012.

[54] H. J. Cordell, "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans," *Human Molecular Genetics,* vol. 11, no. 20, pp. 2463-2468, 2002.

[55] F. J. Poelwijk, V. Krishna and R. Ranganathan, "The context-dependence of mutations: a linkage of formalisms," *PLoS computational biology,* vol. 12, no. 6, p. e1004771, 2016.

[56] Z. R. Sailer and M. J. Harms, "High-order epistasis shapes evolutionary trajectories," *PLoS computational biology,* vol. 13, no. 5, p. e1005541, 2017.

[57] Z. R. Sailer and M. J. Harms, "Detecting high-order epistasis in nonlinear genotype-phenotype maps," *Genetics,* p. 116, 2017.

[58] A. Ay and D. N. Arnosti, "Mathematical modeling of gene expression: a guide for the perplexed biologist," *Critical Reviews in biochemistry and molecular biology,* vol. 46, no. 2, pp. 137-151, 2011.

[59] C. Chaouiya, A. Naldi and D. Thieffry, "Logical modelling of gene regulatory networks with GINsim," in *Bacterial Molecular Networks*, New York NY, Springer, 2012, pp. 463-479.

[60] A. F. Villaverde and J. R. Banga, "Reverse engineering and identification in systems biology: strategies, persepectives and challenges," *Journal of the Royal Society Interface,* vol. 11, no. 91, p. 20130505, 2014.

[61] A. Santos-Zavaleta, M. Sánchez-Pérez, H. Salgado, D. A. Velázquez-Ramírez, S. Gama-Castro, V. H. Tierrafría, S. J. W. Busby, P. Aquino, X. Fang, B. O. Palsson, J. E. Galagan and J. Collado-Vides, "A unified resource for transcriptional regulation in Escherichia coli K-12 incorporating high-

throughput-generated binding data into RegulonDB version 10.0," *BMC biology,* vol. 16, no. 1, p. 91, 2018.

[62] F. D. Urnov, J. C. Miller, Y.-L. Lee, C. M. Beausejour, J. M. Rock, S. Augustus, A. C. Jamieson, M. H. Porteus, P. D. Gregory and M. C. Holmes, "Highly efficient endogenous human gene correction using designed zinc-finger nucleases," *Nature,* vol. 435, no. 7042, p. 646, 2005.

[63] R. Morbitzer, P. Römer, J. Boch and T. Lahaye, "Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors," *Proceedings of the National Academy of Sciences,* vol. 107, no. 50, pp. 21617-21622, 2010.

[64] P. Perez-Pinera, D. D. Kocak, C. M. Vockley, A. F. Adler, A. M. Kabadi, L. R. Polstein, P. I. Thakore, K. A. Glass, D. G. Ousterout, K. W. Leong, F. Guilak, G. E. Crawford, T. E. Reddy and C. A. Gersbach, "RNA-guided gene activation by CRISPR-Cas9-based transcription factors," *Nature Methods,* vol. 10, no. 10, pp. 973-976, 2013.

[65] Y. Jouvenot, V. Ginjala, L. Zhang, P.-Q. Liu, M. Oshimura, A. Feinberg, A. Wolffe, R. Ohlsson and P. Gregory, "Targeted regulation of imprinted genes by synthetic zinc-finger transcription factors," *Gene Therapy,* vol. 10, pp. 513-522, 2003.

[66] N. Corbi, V. Libri and C. Passananti, "Synthetic Zinc Finger Transcription Factors," in *Zinc Finger Proteins*, Boston MA, Springer, 2005, pp. 47-55.

[67] P. Perez-Pinera, D. G. Ousterout, J. M. Brunger, A. M. Farin, K. A. Glass, F. Guilak, G. E. Crawford, A. J. Hartemink and C. A. Gersbach, "Synergistic and tunable human gene activation by combinations of synthetic transcription factors," *Nature Methods,* vol. 10, pp. 239-242, 2013.

[68] A. M. Kabadi and C. A. Gersbach, "Engineering synthetic TALE and CRISPR/Cas9 transciption factors for regulating gene expression," *Methods,* vol. 69, no. 2, pp. 188-197, 2014.

[69] B. J. Hussey and D. R. McMillen, "Programmable T7-based synthetic transcription factors," *Nucleic Acids Research,* vol. 46, no. 18, pp. 9842-9854, 2018.

[70] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero and P. Horvath, "CRISPR provides acquired resistance against viruses in prokaryotes," *Science,* vol. 315, no. 5819, pp. 1709-1712, 2007.

[71] T. Wang, J. J. Wei, D. M. Sabatini and E. S. Lander, "Genetic Screens in Human Cells Using the CRISPR-Cas9 System," *Science,* vol. 343, pp. 80-84, 2014.

[72] O. Parnas, M. Jovanovic, T. M. Eisenhaure, F. Zhang, N. Hacohen and A. Regev, "A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks," *Cell,* vol. 162, pp. 675-686, 2015.

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

[74] S. Konermann, M. D. Brigham, A. E. Trevino, J. Joung, O. O. Abudayyeh, C. Barcena, P. D. Hsu, N. Habib, J. S. Gootenberg, H. Nishimasu, O. Nureki and F. Zhang, "Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex," *Nature,* vol. 517, pp. 583-588, 2015.

[75] O. Shalem, N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, T. S. Mikkelsen, D. Heckl, B. L. Ebert, D. E. Root, J. G. Doench and F. Zhang, "Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells," *Science,* vol. 343, no. 6166, pp. 84-87, 2014.

[76] E. Deltcheva, K. Chylinski, C. M. Sharma, K. Gonzales, Y. Chao, Z. A. Pirzada, M. R. Eckert, J. Vogel and E. Charpentier, "CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III," *Nature,* vol. 471, no. 7340, pp. 602-607, 2011.

[77] M. H. Larson, L. A. Gilbert, X. Wang, W. A. Lim, J. S. Weissman and L. S. Qi, "CRISPR interference (CRISPRi) for sequence-specific control of gene expression," *Nature Protocols,* vol. 8, no. 11, pp. 2180-2196, 2013.

[78] R. Barrangou, "Cas9 Targeting and the CRISPR Revolution," *Science,* vol. 344, no. 6185, pp. 707-708, 2014.

[79] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna and E. Charpentier, "A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity," *Science,* vol. 337, no. 6096, pp. 816-821, 2012.

[80] L. S. Qi, M. H. Larson, L. A. Gilbert, J. A. Doudna, J. S. Weissman, A. P. Arkin and W. A. Lim, "Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression," *Cell,* vol. 152, pp. 1173-1183, 2013.

[81] M. A. Mandegar, N. Huebsch, E. B. Frolov, E. Shin, A. Truong, M. P. Olvera, A. H. Chan, Y. Miyaoka, K. Holmes, C. I. Spencer, L. M. Judge, D. E. Gordon, T. V. Eskildsen, J. E. Villalta, M. A. Horlbeck, L. A. Gilbert, N. J. Krogan, S. P. Sheikh, J. S. Weissman, L. S. Qi, P.-L. So and B. R. Conklin, "CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs," *Cell Stem Cell,* vol. 18, pp. 541-553, 2016.

[82] M. Boettcher and M. T. McManu, "Choosing the Right Tool for the Job: RNAi, TALEN, or CRISPR," *Mol Cell,* vol. 58, no. 4, pp. 575-585, 2015.

[83] Y. H. Sung, I.-J. Baek, D. H. Kim, J. Jeon, J. Lee, K. Lee, D. Jeong, J.-S. Kim and H.-W. Lee, "Knockout mice created by TALEN-mediated gene targeting," *Nature Biotechnology,* vol. 31, pp. 23-24, 2013.

[84] J. Shi, E. Wang, J. P. Milazzo, Z. Wang, J. B. Kinney and C. R. Vakoc, "Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains," *Nature Biotechnology,* vol. 33, pp. 661-667, 2015.

[85] J. P. Shen, D. Zhao, R. Sasik, J. Luebeck, A. Birmingham, A. Bojorquez-Gomez, K. Licon, K. Klepper, D. Pekin, A. N. Beckett, K. S. Sanchez, A. Thomas, C.-C. Kuo, D. Du, A. Roguev, N. E. Lewis, A. N. Chang, J. F. Kreisberg, N. Krogan, L. Qi, T. Ideker and P. Mali, "Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions," *Nature Methods,* vol. 14, no. 6, pp. 573-576, 2017.

[86] D. Žgur-Bertok, "DNA Damage Repair and Bacterial Pathogens," *PLoS pathogens,* vol. 9, no. 11, p. e1003711, 2013.

[87] R. Ceccaldi, B. Rondinelli and A. D. D'Andrea, "Repair Pathway Choices and Consequences at the Double-Strand Break," *Trends in Cell Biology,* vol. 26, no. 1, pp. 52-64, 2016.

[88] D. Du, A. Roguev, D. E. Gordon, M. Chen, S.-H. Chen, M. Shales, J. P. Shen, T. Ideker, P. Mali, L. S. Qi and N. J. Krogan, "Genetic interaction mapping in mammalian cells using CRISPR interference," *Nature Methods,* vol. 14, no. 6, pp. 577-580, 2017.

[89] Y. Fedorov, E. M. Anderson, A. Birmingham, A. Reynolds, J. Karpilow, K. Robinson, D. Leake, W. S. Marshall and A. Khvorova, "Off-target effects by siRNA can induce toxic phenotype," *RNA,* vol. 12, pp. 1188-1196, 2006.

[90] A. L. Jackson and P. S. Linsley, "Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application," *Nature Reviews Drug Discovery,* vol. 9, pp. 57-67, 2010.

[91] A. Birmingham, E. M. Anderson, A. Reynolds, D. Ilsley-Tyree, D. Leake, Y. Fedorov, S. Baskerville, E. Maksimova, K. Robinson, J. Karpilow, W. S. Marshall and A. Khvorova, "3' UTR seed matches, but not overall identity, are associated with RNAi off-targets," *Nature methods,* vol. 3, pp. 199-204, 2006.

[92] B. Zetche, S. E. Volz and F. Zhang, "A Split Cas9 Architecture for Inducible Genome Editing and Transcription Modulation," *Nature Biotechnology,* vol. 33, no. 2, pp. 139-142, 2015.

[93] B. Zetsche, J. S. Gootenberg, O. O. Abudayyeh, I. M. Slaymaker, K. S. Makarova, P. Essletzbichler, S. E. Volz, J. Joung, J. van der Oost, A. Regev, E. V. Koonin and F. Zhang, "Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System," *Cell,* vol. 163, pp. 759-771, 2015.

[94] D. Balboa, J. Weltner, S. Eurola, R. Trokovic, K. Wartiovaara and T. Otonkoski, "Conditionally Stabilized dCas9 Activator for Controlling Gene Expression in Human Cell Reprogramming and Differentiation," *Stem Cell Reports,* vol. 5, pp. 448-159, 2015.

[95] Y. J. Lee, A. Hoynes-O'Connor, M. C. Leong and T. S. Moon, "Programmable control of bacterial gene expression with the combined CRISPR and antisense RNA system," *Nucleic Acids Research,* vol. 44, no. 5, pp. 2462-2473, 2016.

[96] L. A. Gilbert, M. H. Larson, L. Morsut, Z. Liu, G. A. Brar, S. E. Torres, N. Stern-Ginossar, O. Brandman, E. H. Whitehead, J. A. Doudna, W. A. Lim, J. S. Weissman and L. S. Qi, "CRISPR-

Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes," *Cell,* vol. 154, pp. 442-451, 2013.

[97] A. Roguev, D. Talbot, G. L. Negri, M. Shales, G. Cagney, S. Bandyopadhyay, B. Panning and N. J. Krogan, "Quantitative genetic-interaction mapping in mammalian cells," *Nature Methods,* vol. 10, no. 5, pp. 432-437, 2013.

[98] M. E. Tanenbaum, L. A. Gilbert, L. S. Qi, J. S. Weissman and R. D. Vale, "A Protein-Tagging System for Signal Amplification in Gene Expression and Fluorescence Imaging," *Cell,* vol. 159, pp. 635-646, 2014.

[99] Z. Bao, S. Jain, V. Jaroenpuntaruk and H. Zhao, "Orthogonal Genetic Regulation in Human Cells Using Chemically INduced CRISPR/Cas9 Activators," *ACS Synthetic Biology,* vol. 6, pp. 686-693, 2017.

[100] B. Banerjee and R. I. Sherwood, "A CRISPR view of gene regulation," *Current Opinion in Systems Biology,* vol. 1, pp. 1-8, 2017.

[101] L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini and F. Zhang, "Multiplex Genome Engineering Using CRISPR/Cas Systems," *Science,* p. 1231143, 2013.

[102] N. Rajagopal, S. Srinivasan, K. Kooshesh, Y. Guo, M. D. Edwards, B. Banerjee, T. Syed, B. J. M. Emons, D. K. Gifford and R. I. Sherwood, "High-throughput mapping of regulatory DNA," *Nature Biotechnology,* vol. 34, no. 2, pp. 167-174, 2016.

[103] Y. Diao, R. Fang, B. Li, Z. Meng, J. Yu, Y. Qiu, K. C. Lin, H. Huang, T. Liu, R. J. Marina, I. Jung, Y. Shen, K.-L. Guan and B. Ren, "A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells," *Nature Methods,* vol. 14, pp. 629-635, 2017.

[104] C. P. Fulco, M. Munschauer, R. Anyoha, G. Munson, S. R. Grossman and E. M. Perez, "Systematic mapping of functional enhancer-promoter connections with CRISPR interference," *Science,* vol. 354, no. 6313, pp. 769-773, 2016.

[105] T. S. Klann, J. B. Black, M. Chellappan, A. Safi, L. Song, I. B. Hilton, G. E. Crawford, T. E. Reddy and C. A. Gersbach, "CRISPR-Cas9 Epigenome Editing Enables High-Throughput Screening for Functional Regulatory Elements in the Human Genome," *Nature Biotechnology,* vol. 35, pp. 561-568, 2017.

[106] P. Zhang, J.-H. Xia, J. Zhu, P. Gao, Y.-J. Tian, M. Du, Y.-C. Guo, S. Suleman, Q. Zhang, M. Kohli, L. S. Tillmans, S. N. Thibodeau, A. J. French, J. R. Cerhan, L.-D. Wang, G.-H. Wei and L. Wang, "High-throughput screening of prostate cancer risk loci by single nucleotide polymorphisms sequencing," *Nature Communications,* vol. 9, no. 1, p. 2022, 2018.

[107] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, Z. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A.

Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson and J. H. Bielas, "Massively parallel digital transcriptional profiling of single cells," *Nature Communications,* vol. 8, p. 14049, 2017.

[108] B. Adamson, T. M. Norman, M. Jost, M. Y. Cho, J. K. Nunez, Y. Chen, J. E. Villalta, L. A. Gilbert, M. A. Horlbeck, M. Y. Hein, R. A. Pak, A. N. Gray, C. A. Gross, A. Dixit, O. Parnas, A. Regev and J. S. Weissman, "A multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response," *Cell,* vol. 167, pp. 1867-1882, 2016.

[109] T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson and J. Taipale, "Countiing absolute numbers of molecules usiing unique molecular identifiers," *Nature Methods,* vol. 9, no. 1, pp. 72-74, 2012.

[110] S. Xie, A. Cooley, D. Armendarez, P. Zhou and G. C. Hon, "Frequent sgRNA-barcode Recombination in Single-cell Perturbation Assays," *PloS one,* vol. 13, no. 6, p. e0198635, 2018.

[111] P. Datlinger, A. F. Rendeiro, C. Schmidl, T. Krausgruber, P. Traxler, J. Klughammer, L. C. Schuster, A. Kuchler, D. Alpar and C. Bock, "Pooled CRISPR screening with single-cell transcriptome readout," *Nature Methods,* vol. 14, no. 3, pp. 297-301, 2017.

[112] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay and I. Amit, "Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types," *Science,* vol. 343, pp. 776-779, 2014.

[113] M. W. Gander, J. D. Vrana, W. E. Voje, J. M. Carothers and E. Klavins, "Digital logic circuits in yeast with CRISPR-dCas9 NOR gates," *Nature Communications,* vol. 8, p. 15459, 2017.

[114] A. A. Nielsen and C. A. Voigt, "Multi-input CRISPR/Cas genetic circuits that interface host regulatory networks," *Molecular Systems Biology,* vol. 10, no. 11, p. 763, 2014.

[115] R. H. Katz and G. Borriello, Contemporary logic design, Upper Saddle River, New Jersey: Pearson Prentice Hall, 2005.

[116] A. Tamsir, J. J. Tabor and C. A. Voigt, "Robust multicellular computing using genetically encoded NOR gates and chemical 'wires'," *Nature,* vol. 469, pp. 212-215, 2011.

[117] B. C. Stanton, A. A. K. Nielsen, A. Tamsir, K. Clancy, T. Peterson and C. A. Voigt, "Genomic mining of prokaryotic repressors for orthogonal logic gates," *Nature Chemical Biology,* vol. 10, pp. 99-105, 2014.

[118] J. G. Zalatan, M. E. Lee, R. Almeida, L. A. Gilbert, E. H. Whitehead, M. La Russa, J. C. Tsai, J. S. Weissman, J. E. Dueber, L. S. Qi and W. A. Lim, "Engineering Complex Synthetic Transcriptional Programs with CRISPR RNA Scaffolds," *Cell,* vol. 160, pp. 339-350, 2015.

[119] Y. Liu, Y. Zeng, L. Liu, C. Zhuang, X. Fu, W. Huang and Z. Cai, "Synthesizing AND gate genetic circuits based on CRISPR-Cas9 for identification of bladder cancer cells," *Nature Communications,* vol. 5, p. 5393, 2014.

[120] L. Nissim, S. D. Perli, A. Fridkin, P. Perez-Pinera and T. K. Lu, "Multiplexed and Programmable Regulation of Gene Networks with an Integrated RNA and CRISPR/Cas Toolkit in Human Cells," *Molecular Cell,* vol. 54, pp. 698-710, 2014.

[121] T. Hoshino, "Violacein and related tryptophan metabolites produced by Chromobacterium violaceum: biosynthetic mechanism and pathway for construction of violacein core," *Applied Microbiology and Biotechnology,* vol. 91, no. 6, pp. 1463-1475, 2011.

[122] B. Zetsche, M. Heidenreich, P. Mohanraju, I. Fedorova, J. Kneppers, E. M. DeGennaro, N. Winblad, S. R. Choudhury, O. O. Abudayyeh, J. S. Gootenberg, W. Y. Wu, D. A. Scott, K. Severinov, J. van der Oost and F. Zhang, "Multiplex gene editing by CRISPR-Cpf1 through autonomous processing of a single crRNA array," *Nature Biotechnology,* vol. 35, no. 1, pp. 31-34, 2017.

[123] P.-A. Gros, H. Le Nagard and O. Tenaillon, "The evolution of epistasis and its links with genetic robustness, complexity and drift in a phenotypic model of adaptation," *Genetics,* 2009.

[124] F. J. Poelwijik, S. Tănase-Nicola, D. J. Kiviet and S. J. Tans, "Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes," *Journal of Theoretical Biology,* vol. 272, no. 1, p. 141, 2011.

[125] K. Chiotti, D. J. Kvitek, K. Schmidt, G. Koniges, K. Schwartz, E. A. Donckels, F. Rosenzweig and G. Sherlock, "The valley of death: reciprocal sign epistasis contrains adaptive trajectories in a constant, nutrient limiting environment," *Genomics,* vol. 104, no. 6, pp. 431-437, 2014.

[126] D. S.-W. Ow, P. M. Nissom, R. Philp, S. K.-W. Oh and M. G.-S. Yap, "Global transcriptional analysis of metabolic burden due to plasmid maintenance in Escherichia coli DH5α during batch fermentation," *Enzyme and Microbial Technology,* vol. 39, no. 3, pp. 391-398, 2006.

[127] H. Dong, L. Nilsson and C. G. Kurland, "Gratuitous overexpression of genes in Escherichia coli leads to growth inhibition and ribosome destruction," *Journal of bacteriology,* vol. 177, no. 6, pp. 1497-1504, 1995.

[128] G. Claes, S. Govindarajan and J. Minshull, "Codon bias and heterologous protein expression," *Trends in biotechnology,* vol. 22, no. 7, pp. 346-353, 2004.

[129] K. Zhao, M. Liu and R. R. Burgess, "Adaptation in bacterial flagellar and motility systems: from regulon members to 'foraging'-like behavior in E. coli," *Nucleic Acids Research,* vol. 35, no. 13, pp. 4441-4452, 2007.

[130] E. Martínez-García, P. I. Nikel, M. Chavarría and V. de Lorenzo, "The metabolic cost of flagellar motion in Pseudomonas putida KT 2440," *Environmental microbiology,* vol. 16, no. 1, pp. 291-303, 2014.

[131] H. Sheftel, O. Shoval, A. Mayo and U. Alon, "The geometry of the Pareto front in biological phenotype space," *Ecology and Evolution,* vol. 3, no. 6, pp. 1471-1483, 2013.

[132] J. J. Faith, M. E. Discoll, V. A. Fusaro, E. J. Cosgrove, B. Hayete, F. S. Juhn, S. J. Schneider and T. S. Gardner, "Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata," *Nucleic Acids Research,* pp. D866-D870, 2008.

[133] R. A. Lease, M. E. Cusick and M. Belfort, "Riboregulation in Escherichia coli: DsrA RNA acts by RNA:RNA interactions at multiple loci," *Proceedings of the National Academy of Sciences,* vol. 95, no. 21, pp. 12456-12461, 1998.

[134] F. Karginov and G. J. Hannon, "The CRISPR system: small RNA-guided defense in bacteria and archaea," *Molecular Cell,* vol. 37, no. 1, pp. 7-19, 2010.

[135] L. Cui and D. Bikard, "Consequences of Cas9 cleavage in the chromosome of Escherichia coli," *Nucleic Acids Research,* vol. 44, no. 9, pp. 4243-4251, 2016.

[136] A. S. L. Wong, G. C. G. Choi, C. H. Cui, G. Pregernig, P. Milani, M. Adam, S. D. Perli, S. W. Kazer, A. Gaillard, M. Hermann, A. K. Shalek, E. Fraenkel and T. K. Lu, "Multiplexed barcoded CRISPR-Cas9 screening enabled by CombiGEM," *Proceedings of the National Academy of Sciences,* p. 201517883, 2016.

[137] F. J. Poelwijk and R. Ranganathan, "The relationship between alignment covariance and background-averaged epistasis," *arXiv preprint,* vol. 1703, p. 10996, 2017.

[138] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens and Y. Gilad, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Research,* 2008.

[139] Z. Wang, M. Gerstein and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature reviews genetics,* vol. 10, pp. 57-63, 2009.

[140] X. Fu, N. Fu, S. Guo, Z. Yan, Y. Xu, H. Hu, C. Menzel, W. Chen, Y. Li, R. Zeng and P. Khaitovich, "Estimating accuracy of RNA-Seq and microarrays with proteomics," *BioMed Central Genomics,* vol. 10, no. 1, p. 161, 2009.

[141] J. H. Malone and B. Oliver, "Microarrays, deep sequencing and the true measure of the transcriptome," *BioMed Central Biology,* vol. 9, no. 1, p. 34, 2011.

[142] X. Dai, M. Zhu, M. Warren, R. Balakrishnan, V. Patsalo, H. Okano, J. R. Williamson, K. Fredrick, Y.-P. Wang and T. Hwa, "Reduction of translating ribosomes enables Escherichia coli to maintain elongation rates during slow growth," *Nature microbiology,* vol. 2, no. 2, p. 16231, 2016.

[143] M. Basan, S. Hui, H. Okano, Z. Zhang, Y. Shen, J. R. Williamson and T. Hwa, "Overflow metabolism in Escherichia coli results from efficient proteome allocation," *Nature,* vol. 528, no. 7580, p. 99, 2015.

[144] C. Furlanello, M. Serafini, S. Merler and G. Jurman, "Entropy-based gene ranking without selection bias for the predictive classification of microarray data," *BMC Bioinformatics,* vol. 4, no. 1, p. 54, 2003.

[145] R. Tibshirani, G. Walther and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B,* vol. 63, no. 2, pp. 411-423, 2001.

[146] J. Oyelade, I. Isewon, F. Oladipupo, O. Aromolaran, E. Uwoghiren, F. Ameh, M. Achas and E. Adebiyi, "Clustering algorithms: Their application to gene expression data," *Bioinformatics and Biology insights,* vol. 10, pp. BBI-S38316, 2016.

[147] J. Luo, M. Schumacher, A. Scherer, D. Sanoudou, D. Megherbi, T. Davison, T. Shi, W. Tong, L. Shi, H. Hong, C. Zhao, F. Elloumi, W. Shi, R. Thomas, S. Lin, G. Tillinghast, G. Liu, Y. Zhou, D. Herman, Y. Li, Y. Deng, H. Fang, P. Bushel, M. Woods and J. Zhang, "A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data," *The Parmacogenomics Journal,* vol. 10, pp. 278-291, 2010.

[148] C. Kahramanoglou, A. S. N. Seshasayee, A. I. Prieto, D. Ibberson, S. Schmidt, J. Zimmermann, V. Benes, G. M. Fraser and N. M. Luscombe, "Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli," *Nucleic Acids Research,* vol. 39, no. 6, pp. 2073-2091, 2011.

[149] J.-C. Baret, O. J. Miller, V. Taly, M. Ryckelynck, A. El-Harrak, L. Frenz, C. Rick, M. L. Samuels, J. B. Hutchison, J. J. Agresti, D. R. Link, D. A. Weitz and A. D. Griffiths, "Fluorescence-activated droplet sorting (FADS): efficient microfluidic cell sorting based on enzymatic activity," *Lab on a Chip,* vol. 9, no. 13, pp. 1850-1858, 2009.

[150] E. Tacconelli, E. Carrara, A. Savoldi, S. Harbarth, M. Mendelson, D. L. Monnet, C. Pulcini, G. Kahlmeter, J. Kluytmans, Y. Carmeli, M. Ouellette, K. Outterson, J. Patel, M. Cavaleri, E. M. Cox, C. R. Houchens, M. L. Grayson, P. Hansen, N. Singh, U. Theuretzbacher and N. Magrini, "Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis," *The Lancet,* vol. 18, no. 3, pp. 318-327, 2018.

[151] A. Zipperer, M. C. Konnerth, C. Laux, A. Berscheid, D. Janek, C. Weidenmaier, M. Burian, N. A. Schilling, C. Slavetinsky, M. Marschal, M. Willmann, H. Kalbacher, B. Schittek, H. Brötz-Oesterhelt, S. Grond, A. Peschel and B. Krismer, "Human commensals producing a novel antibiotic impair pathogen colonization," *Nature,* vol. 535, pp. 511-516, 2016.

[152] L. L. Ling, T. Schneider, A. J. Peoples, A. L. Spoering, I. Engels, B. P. Conlon, A. Mueller, T. F. Schäberle, D. E. Hughes, S. Epstein, M. Jones, L. Lazarides, V. A. Steadman, D. R. Cohen, C. R. Felix, K. A. Fetterman, W. P. Millett, A. G. Nitti, A. M. Zullo, C. Chen and K. Lewis, "A new antibiotic kills pathogens without detectable resistance," *Nature,* vol. 517, pp. 455-459, 2015.

[153] K. Sciarretta, J.-A. Røttingen, A. Opalska, A. J. Van Hengel and J. Larsen, "Economic Incentives for antibacterial drug development: literature review and considerations from the transatlantic task

force on antimicrobial resistance," *Clinical Infectious Diseases,* vol. 63, no. 11, pp. 1470-1474, 2016.

[154] A. Mullard, "New Drugs Cost US$2.6 billion to develop," *Nature reviews drug discovery,* vol. 13, no. 12, p. 877, 2014.

[155] A. R. Brochado, A. Telzerow, J. Bobonis, M. Banzhaf, A. Mateus, J. Selkrig, E. Huth, S. Bassler, J. Z. Beas, M. Zietek, N. Ng, S. Foerster, B. Ezraty, B. Py, F. Barras, M. M. Savitski, P. Bork, S. Göttig and A. Typas, "Species-specific activity of antibacterial drug combinations," *Nature,* vol. 559, pp. 259-263, 2018.

[156] A. Kulesa, J. Kehe, J. E. Hurtado, P. Tawde and P. C. Blainey, "Combinatorial drug discovery in nanoliter droplets," *Proceedings of the National Academy of Sciences,* p. 201802233, 2018.

[157] F. Eduati, R. Utharala, D. Madhavan, U. P. Neumann, T. Longerich, T. Cramer, J. Saez-Rodriguez and C. A. Merten, "A microfluidics platform for combinatorial drug screening on cancer biopsies," *Nature Communications,* vol. 9, no. 1, p. 2434, 2018.

[158] K. A. Wetterstrand, "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)," National Human Genome Research Institute, NIH, 25 April 2018. [Online]. Available: https://www.genome.gov/27541954/dna-sequencing-costs-data/. [Accessed 28 August 2018].

[159] R. Zilionis, J. Nainys, A. Veres, V. Savova, D. Zemmour, A. M. Klein and L. Mazutis, "Single-Cell barcoding and sequencing using droplet microfluidics," *Nature protocols,* vol. 12, no. 1, p. 44, 2017.

[160] Y. Xin, J. Kim, M. Ni, Y. Wei, H. Okamoto, J. Lee, C. Adler, K. Cavino, A. J. Murphy, G. D. Yancopoulos, H. C. Lin and J. Gromada, "Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells," *Proceedings of the National Academy of Sciences,* vol. 113, no. 12, pp. 3293-3298, 2016.

[161] T. M. Gierahn, M. H. Wadsworth, T. K. Hughes II, B. D. Bryson, A. Butler, R. Satija, S. Fortune, J. C. Love and A. K. Shalek, "Seq-Well: A Portable, Low-Cost Platform for High-Throughput Single-Cell RNA-Seq for Low-Input Samples," *Nature methods,* vol. 14, no. 4, p. 395, 2017.

[162] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulous, K. Bealer and T. L. Madden, "BLAST+: architecture and applications," *BMC Bioinformatics,* vol. 10, no. 1, p. 421, 2009.

[163] H. B. Jensen and K. Kleppe, "Effect of Ionic Strength, pH, Amines and Divalent Cations on the Lytic Activity of T4 Lysozyme," *European journal of biochemistry ,* vol. 28, no. 1, pp. 116-122, 1972.

[164] A. Meyerhans, J.-P. Vartanian and S. Wain-Hobson, "DNA recombination during PCR," *Nucleic acids research,* vol. 18, no. 7, pp. 1687-1691, 1990.

[165] S. Chong, C. Chen, H. Ge and X. S. Xie, "Mechanism of transcriptional bursting in bacteria," *Cell,* vol. 158, no. 2, pp. 314-326, 2014.

[166] K. Bush, P. Courvalin, G. Dantas, J. Davies, B. Eisenstein, P. Huovinen, G. A. Jacoby, R. Kishony, B. N. Kreiswirth, E. Kutter, S. A. Lerner, S. Levy, K. Lewis, O. Lomovskaya, J. H. Miller, S. Mobashery, L. J. V. Piddock, S. Projan, C. M. Thomas, A. Tomasz, P. M. Tulkens, T. R. Walsh, J. D. Watson, J. Witkowski, W. Witte, G. Wright, P. Yeh and H. I. Zgurskaya, "Tackling antibiotic resistance," *Nature reviews,* vol. 9, pp. 894-896, 2011.

[167] L. R. Mulcahy, J. L. Burns, S. Lory and K. Lewis, "Emergence of Pseudomonas aeruginosa strains producing high levels of persister cells in patients with cystic fibrosis," *Journal of bacteriology,* vol. 192, no. 23, pp. 6191-6199, 2010.

[168] J. W. Bioger, "Treatment of staphylococcal infections with penicillin by intermittent sterilization," *Lancet,* vol. 14, p. 497, 1944.

[169] K. Lewis, "Persister cells," *Annual review of microbiology,* vol. 64, pp. 357-372, 2010.

[170] S. Helaine and E. Kugelberg, "Bacterial persisters: formation, eradication, and experimental systems," *Trends in microbiology,* vol. 22, no. 7, pp. 417-424, 2014.

[171] C. I. Kint, N. Verstraeten, M. Fauvart and J. Michiels, "New-found fundamentals of bacterial persistence," *Trends in microbiology,* vol. 20, no. 12, pp. 577-585, 2012.

[172] Z. Baharoglu and D. Mazel, "Vibrio cholerae triggers SOS and mutagenesis in response to a wide range of antibiotics: a route towards multiresistance," *Antimicrobial agents and chemotherapy,* vol. 55, no. 5, pp. 2438-2441, 2011.

[173] Z. Baharoglu, E. Krin and D. Mazel, "RpoS plays a central role in the SOS induction by sub-lethal aminoglycoside concentraions in Vibrio cholerae," *PLoS genetics,* vol. 9, no. 4, p. e1003421, 2013.

[174] I. Keren, N. Kaldalu, A. Spoering, Y. Wang and K. Lewis, "Persister cells and tolerance to antimicrobials," *FEMS microbiology letters,* vol. 230, no. 1, pp. 13-18, 2004.

[175] D. Shah, Z. Zhang, A. B. Khodursky, N. Kaldalu, K. Kurg and K. Lewis , "Persisters: a distinct physiological state of E. coli," *BMC microbiology,* vol. 6, no. 1, p. 53, 2006.

[176] Qiagen, "FAQ: How much RNA does a typical mammalian cell contain?," Qiagen, [Online]. Available: https://www.qiagen.com/fr/resources/faq?id=06a192c2-e72d-42e8-9b40-3171e1eb4cb8&lang=en. [Accessed 27 08 2018].

[177] R. Milo and R. Phillips, "Cell biology by the numbers," Bionumbers, [Online]. Available: http://book.bionumbers.org/how-many-mrnas-are-in-a-cell/. [Accessed 01 09 2018].

[178] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model," *Nature Methods,* vol. 12, no. 10, pp. 931-934, 2015.

[179] M. Pirkl and N. Beerenwinkel, "Single cell network analysis with a mixture of Nested Effects Models," *bioRxiv,* p. 258202, 2018.