SLIM: Stochastic Lineage-based Iterative Minimization

Thesis by

Mikel Lipschitz

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2026

(Defended August 14th, 2025)

Mikel Lipschitz
ORCID: 0000-0002-5764-1648

ACKNOWLEDGEMENTS

It truly does take a village to raise a graduate student (thanks for stealing this great line and using it first, Matteo!). I have so many people to thank for both directly and indirectly helping me reach this point in my career. I'll start by thanking my former mentor, Dr. Scott Rodig for his unwavering patience, mentorship and support. When I started in Scott's lab I could barely even hold a pipette. By the time I was finished I was listed as co-author on at least a dozen publications. I attribute very little of that success to my skills as a researcher and almost all it to Scott's willingness to simultaneously indulge my ~~insanity~~ creativity while reigning me in, keeping me on track and rewarding me with such wonderful opportunities that I almost certainly didn't deserve. Being included in high-stakes affairs such as publications, professional meetings and conferences, as a humble research technician, really gave me confidence and fueled my desire to become a science lifer.

While I would not have gotten anywhere without Scott's tutelage, I quite literally never would have made it to Caltech without the assistance of Dr. David Van Valen. Dave is the reason why I applied to Caltech in the first place and Dave is the reason why I landed an interview, no question about it. We only worked together for a criminally short period of time, but I will always be grateful for the opportunity he afforded me.

During my time at Caltech I have met so many people that have impacted my life, both personally and professionally, for the better. The great work of Kaihang and so many members of the Wang Lab provided the foundation for much of what you will see in this thesis. Thorough work is a hallmark of high-level science, you will find it in every single lab at this and other top-tier institutions. Mentorship and selflessness are, however, less prevalent. These two characteristics are embodied by two members of the Wang Lab in particular: Dr. Russell Swift and Dr. Charles Sanfiorenzo. Russell not only guided me through a lot of the science that I worked on, but also through the sociological and science adjacent aspects of lab life. He was always open, honest, available and fair, rare qualities indeed. He generously dedicated much time and effort into making sure everybody had a safe place to turn to, for any manner, science or otherwise. He did all this in way that seemed almost unintentional. Maybe it was just instinctual for him, just the kind of person that he is. Russell left the lab a couple of years ago to pursue other opportunities; his presence continues to be sorely missed.

Charles is about a decade younger than me and began graduate school just a year before me. Not exactly the expected characteristics for a mentor. From my first day in the lab, it was clear that Charles is an incredibly talented scientist. We have worked closely on so many projects, both within the context of the lab and outside of it. In fact, there is almost nothing done in the lab that doesn't have Charles imprint on it. He works longer and harder than anyone I have met during my

research career, or really any other aspect of life. This alone would qualify Charles as impressive; however, it is his tremendous growth as a mentor that has impressed me the most. Charles has selflessly managed numerous projects, ensuring that fellow graduate students in the lab remain on track all while navigating his own graduate student experience. He creates opportunities for others and ensures people get the credit they deserve. I am grateful that I have had the opportunity to work with Charles, I am grateful for all the thought and effort that he has put into the projects that we have worked on together and I am grateful for the fantastic insights and knowledge that he has shared with me regarding my own work.

I want to thank my fantastic committee: Dr. Mikhail Shapiro, Dr. Bruce Hay and Dr. Rob Phillips. Looking back, I regret not having more committee meetings. It was always such a treat for me to have such titans of science all come together to hear me talk about my work. The attention and insights that the committee afforded me, especially during times that I believed that nobody really cared about what I was working on, gave me so much confidence. Bruce and Rob, in particular, have each gone far beyond the expectations of a committee member. Through several meetings, Rob helped me to navigate the final stages of my graduate student experience and prepare for my future. During these meetings, so informal and friendly in nature, Rob always seemed say something that would boost my enthusiasm for my work and the prospects for my future. These meetings did and will continue to carry great meaning for me.

Through the final months of my tenure as a graduate student Bruce has played a pivotal role in helping me carve out a path for my future. He has facilitated introductions, included me in meetings and dedicated substantial time to helping me explore and refine ideas. Because of this, I may actually be able to pursue my passions immediately and not just bide time until something falls into my lap. His time, advice and nsights have been invaluable to me. I can only hope this continues long into the future.

To my friends; from the Flubber Ducks, to the BMB Bombers to Sotto Troie to the masterful geniuses behind Party Wine, wouldn't it have been great if we didn't have all of these burdensome research obligations always popping up and ruining our good time? There are so many people amongst each of these groups and some across all of them that have time and time again lifted me out of the depths of despair and reminded me that fun and enjoyment and happiness are all things that I can have, even as a lowly graduate student. Tom, Andy, Matteo, Linlin, Khalil, Victoria, Laura, Drew, Jamie, Victor, Arjuna, Al (a real Bombers legend) and so many others, you all gave me the kind of community that turns a place you move to for grad school into a home that you will never forget.

To my family as a whole; your support and encouragement and love has made all of this possible. It takes a certain degree of selfishness to become a scientist. I have

missed so much over the course of my graduate studies, yet your support has never wavered. To Sari and Elvin; you guys routinely house me, feed me, clothe me and entertain me, you keep me connected and anchored and never ask for anything in return. I hope one day I can give back to you even just a fraction of what you have given me. To my nephews, Miles and Theo; I have been in school as long as you have known me. (As a 40 year old student, I guess I have been in school as long as most people have known me). I hope you can take my experiences as an example that you if you want something, if you want to be something, you can just go out there and be it. This thesis is proof that you can achieve whatever the hell you want and I hope nobody ever convinces you otherwise. I can't wait to watch you grow and see the people you become. I'll be there every step of the way with advice and insight, whether you like it or not.

To my parents; the people that quite literally made all of this possible. I can't even imagine what a ride it has been, having a son like me. I don't know when I morphed into a person that refuses to do anything in any way but my own. I don't know how you have managed to stand by me and support me through the many (many!) head-scratching decisions I have made in my life. Mostly though, I am glad that I don't know what it's like to go a day without knowing that I am loved and supported. That has given me the strength and conviction to not just complete this degree, but to pursue a career in science in general. You two are extraordinary and selfless and, honestly, way more fun and interesting than young me ever gave you credit for.

Finally, to Nikki; my partner, my best friend and my forever adventure buddy. The energy you bring into every room, every conversation, fuels me. The rigor and passion that you bring to your own work never ceases to amaze and inspire me. There is so much I can say here but I will keep it simple, with just the most important point. Throughout all of this time, during my thesis work, you never stopped reminding me that what I was doing was interesting and important, that it mattered. No matter how many times I tried to convince you of the contrary, you refused to budge. This meant EVERYTHING to me. It is only because of your strength and wisdom and love that I could reach this point.

I dedicate this work to you, Nikki.

"Imagine what you'll know tomorrow"

Tommy Lee Jones as 'K'

Men in Black

ABSTRACT

The bacterial genomes we encounter today have been shaped by billions of years of genome altering events which involve rewriting, addition, and removal of genomic elements. The resulting product is a complex network of interactions composed of elements which are defined by contextual necessity. The elucidation of a minimal set of elements, comprised of just those essential for sustaining life, has long been sought after. This set, or minimal genome, has been proposed to be a representation for the foundation of life itself. Genome minimization, the pursuit of this foundation, is a process by which genomic segments deemed unnecessary, or non-essential, in an environmental context dependent manner are identified and removed, leaving only DNA that provides the cell with the resources and processes it needs to stay alive and reproduce. Numerous genome minimization efforts have been undertaken previously. However, each of these studies has resulted in the generation of a single genome-reduced strain derived from a single wild-type bacteria in a single environment. While these methods have shown a great deal of promise in their ability to identify foundational genomic pieces in this extremely narrow context, they lack the throughput and generalizability to identify foundational pieces of all bacterial life.

Building upon prior genome minimization efforts, we developed SLIM (Stochastic Lineage-based Iterative Minimization), a modular genome reduction system designed for unbiased DNA removal, high-throughput parallelization and cross-species compatibility. In this study we utilize SLIM to generate a library of ten genome-reduced *E. coli* strains. We then rigorously interrogate the library to identify patterns in deleted segments. We assess the effects that these deletions have on remaining genomic components and explore how these effects can result in substantial fitness changes in different environments. Finally, we demonstrate the modularity of SLIM by generating two additional libraries of genome-reduced strains from two phylogenetically distinct parent bacteria, *S. flexneri* and *P. putida*. This work highlights the power and promise of generating diverse libraries of

genome-reduced strains; substantially expanding the number of minimized genomes that can be achieved, while simultaneously reducing the time to generation.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Number *Page*

PROLOGUE

The genome is a paradox; a nearly invisible ecosystem so robust it orchestrates every trait of every living thing, yet so fragile that a single change can spell certain death. Genes, the protein-coding elements of the genome, encode a dazzling array of functions and features. Regulatory and other non-coding regions determine when, where, and how strongly these genetic instructions are utilized. Together, they generate the stunning diversity of life on Earth: the riot of colors in flora and fauna, the fragrance of flowers, the sweetness of fruit. From the towering giant sequoia to the humble dwarf willow, it is the genome that shapes every form of life, and it is our own genomes that govern how we perceive and interact with them.

Every living cell contains a genome. Across different cell types within an individual, such as lymphocytes and neurons, the genome is nearly identical. The differences arise from context-dependent gene expression. Even between two humans, genomic sequences are more than 99% identical[1]. Between a human cell and a mouse cell, gene-to-gene similarity is around 80%[2]; with a banana, up to 25%[3]. How can things so visually and functionally distinct share so much genetic material? This shared genetic foundation reflects life's common roots, fundamental elements that transcend species. Yet, the smallest deviation in the genome can lead to profound effects: a single mutation might turn dark hair blonde[4] or trigger a genetic disease[5–8].

The genome is not life itself, but it is the blueprint, the recipe, and the regulator of life.

Chapter 1

AIN'T NOTHIN BUT A GENOME THANG, BABY

Shaping of genomes, through the years

The bacterial genomes we encounter today are the products of billions of years of evolution. Throughout this time they have been shaped by a continuous onslaught of genome altering events which involve rewriting, addition and removal of genomic elements[9]. Natural selection plays a key role in determining the fate of genomic alterations, whether they are retained, propagated, or purged.

When a genome altering event occurs, if the event provides a fitness advantage it will likely propagate through future generations of a population and be maintained. As populations expand and colonize new environments, cells are likely to experience additional genome altering events. Genes are gained through duplication, horizontal transfer, and adaptive innovation via rewriting, but also lost through drift and other modes of streamlining. Many bacterial genomes contain a combination of universally conserved functions, lineage-specific adaptations, and genomic "baggage" accumulated over time. This baggage includes mobile elements, cryptic prophages, pseudogenes, and other relics of past conflicts and transitions. Evolution rarely produces streamlined genomes optimized for engineering, it produces functional patchworks shaped by local fitness constraints. As a result, what is essential in one organism under one set of conditions may be dispensable in another. The essentiality of a gene is path-dependent: shaped by the history and current configuration of the genome in which it resides.

While natural selection works to shape genomes over the course of generations, some DNA modulation events occur during the lifetime of a single cell. The major modes of DNA acquisition are viral infection and horizontal gene transfer through conjugation[10]. In many viral infections, the viral genome is often integrated into the

host genome[11,12]. Any host that is able to stave off the inevitable lysis event due to virion reproduction may retain part of or the entire viral genome embedded in its own[13]. Horizontal gene transfer, often mediated by conjugation, plays a major role in a bacteria's accumulation of genomic elements both within and across generations. Advantageous DNA can propagate through populations quickly. This DNA often comes in the form of mobile genetic elements (MGE) such as transposons and insertion sequences (IS). MGE are the underlying force behind DNA transfer and amplification. These elements are typically composed of segments of DNA containing MGE effector recognition sites and cargo. There two major distinct mechanisms of action amongst MGE: copy-and-paste MGE makes copies of themselves and the copy inserts into a different location in either the same DNA molecule or a different DNA molecule[14], whereas cut-and-paste MGE excise themselves out of their original location and insert into a new location in either the same or a different DNA molecule[14]. MGE insertions can alter the expression of their cargo genes and of neighboring host genes. The exact movements of any specific MGE amongst a population of cells carries a degree of randomness. The ability to transit DNA molecules allows MGE to position themselves for dramatic amplification and/or cell-cell transfer. MGE activity can be both advantageous and detrimental, though it is the advantageous movements which propagate through generations.

DNA removal or loss events can be just as prevalent as DNA addition events[15]. Genomic segments can be lost due to homologous recombination in combination with or independent of deficient replication. Deficient replication events typically involve replication fork stalling or collisions, leading to chromosomal fragmentation[16], or replication misalignment[17]. DNA repair mechanisms involving homologous recombination, following chromosomal fragmentation, can result in DNA loss that is propagated through generations; meaning progeny may contain genomes which are missing segments of their parent genomes. In standard homologous recombination-mediated deletion events, flanking redundant

sequences can recombine and excise the intervening segment. Replication misalignment events can result in the skipping of intermediate DNA flanked by repetitive sequences. As with genomic alterations through DNA addition, DNA removal events can be either advantageous or detrimental. Advantageous DNA removal events typically involve the removal of burdensome DNA, sometimes following mutation events[15].

DNA rewriting, or mutational, events parallel DNA addition and removal events. Mutations, which can manifest as single nucleotide polymorphisms or insertions/deletions, can result in the activation, deactivation or repurposing of specific gene products. Through natural selection, advantageous genome altering events are maintained while detrimental events are lost. Together, DNA acquisition, loss and rewriting create a dynamic genomic equilibrium shaped by environmental pressures.

While DNA acquisitions and losses occur during the span of a single generation, it is through natural selection over the course of many generations that these shifts in genomic content are stabilized. Genomes today contain remnants of billions of years of these events. During this time, many genomic sequences have had the opportunity to coevolve together, creating a complex landscape of interaction networks in which genes can work together, against each other or instead of each other. The magnitude of these interaction networks presents a substantial barrier to completely understanding the full functionality of all genes and the relationships between them. Despite the decades of research dedicated to expanding the understanding of the functions of and relationships between genes in *E. coli*, the most studied organism in existence, the functions of only 65% of its genes have been elucidated[18]. Even less is known about the regulatory mechanisms of all genes. The inability to fully elucidate the regulation, function and relationships of all genes in *E. coli* is a direct result of the complexity that the genome has achieved over billions of years. The intricate set of gene interaction networks makes it difficult to

identify functions and interactions for all genes. This is not only true for *E. coli*, but is exacerbated for other, less annotated, organisms.

Overview of genome minimization, goals and methods

Genome minimization is the process by which genomic segments deemed unnecessary, or non-essential, in an environmental context dependent manner are identified and removed, leaving only DNA that provides the cell with the resources and processes it needs to stay alive and reproduce. Specifically, genome minimization seeks to remove redundant or non-essential elements to simplify genetic architecture as part of a general complexity reduction process.

In the process of reducing complexity, many genomic segments and processes for which they encode are removed, freeing up energy to be used for remaining processes. As such, genome reduction is sometimes used to generate bacteria which are more functionally optimized to complete highly specific tasks, such as recombinant protein production[19–22].

Due to the intricate network of interactions apparent across all components of the genome, it can be difficult to fully elucidate regulation and function of individual genes and gene products and interactions between multiple gene products in wild-type genomes. The difficulty of this task is often amplified by conditional and/or low-level expression of some genes. The removal of non-essential DNA reduces complexity of the network, allowing for a more focused exploration of unknown functions, interactions and regulatory mechanisms amongst incompletely annotated genes. As a result, reduced genomes offer a unique platform to dissect fundamental biological principles by identifying core genomic elements, or building blocks, essential to life. Ultimately, the goal of many genome minimization efforts is to construct a chassis composed of only the most foundational building blocks. Such a chassis could serve as a clean slate for biological engineering, offering a modular platform upon which synthetic components can be added without unintended interference or complexity.

The result of genome minimization is highly dependent on the environment in which the process takes place. As previously mentioned, cells are equipped with genomic elements allowing them to survive and even thrive in a wide range of environments. This diversity of features is important for the survival of a species through generations in vast environments. But, in any one environment, many of those genomic elements become unnecessary for survival. Since the process of genome minimization takes place in a single laboratory environment, those same unnecessary genomic elements can be removed with no observable cost to cellular viability under laboratory conditions. There are two factors that arise as a result of this relationship between environment and genomic necessity. First, conducting the genome reduction process in different environments will result in distinct reduced genomes. Second, conducting the genome reduction process in one environment will dramatically alter the behavior of the genome reduced strain in a different environment. These factors must be considered when evaluating the results of any genome reduction process.

There are two general strategies for genome reduction: bottom-up minimization and top-down minimization.

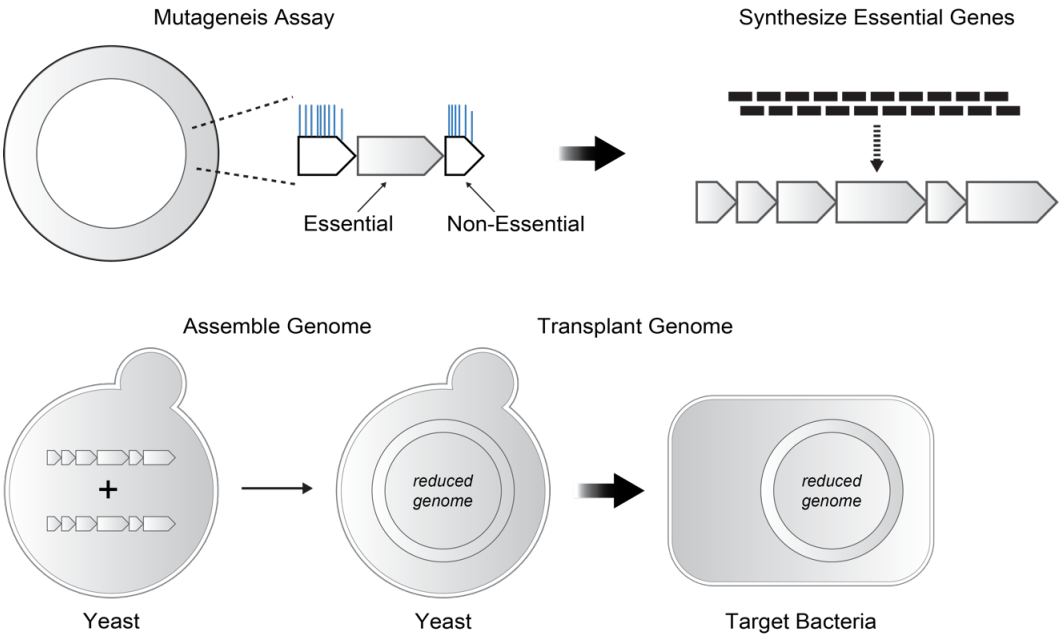

Figure 1.1: Schematic representation of bottom-up minimization
A typical bottom-up minimization scheme follows the following steps: (1) Delineate essential from non-essential portions of the genome. (2) Design minimal genome based on information from step 1 and synthesize essential into DNA segments. (3) Assemble DNA segments in yeast and then (4) transplant into the target bacterium.

The main premise of bottom-up minimization is to generate a defined, minimal set of genes, synthesize the set into a contiguous DNA molecule and then transplant the synthesized minimal set of genes into a recipient cell (**Figure 1.1**). If the minimal set contains all requirements to sustain life, the cell will survive; if not, the cell will die. It is important to note that, even in the event of cell survival, there is no guarantee that the genome has been fully minimized. The typical minimal set of genes used in any bottom-up minimization study is usually derived from various sources, including essentiality and cellular fitness studies[23]. Essentiality studies often employ a mutagenesis mechanism to knockout a single gene per cell in a pool of cells, producing a diverse assortment of knockouts across the entire pool[24,25]. The output of these studies can be considered a first-order essentiality classification of all genes. Meaning, these studies assess essentiality in the context of an otherwise unmodified genome, without accounting for combinatorial gene interactions. This limitation can potentially result in a failure to identify synthetic lethal pairs and emerging essential or quasi-essential genes as the minimization process progresses, which can only be uncovered through iterative or combinatorial approaches. The failure to correctly identify genes in these categories can result in a stalling of the genome reduction process. As a result of this all-or-nothing approach, bottom-up minimization studies often get stuck in the design-build-test cycle.

In 2016, generation of the first fully synthetic reduced bacterial genome was reported[26]. In this study a *Mycoplasma mycoides* strain, JCVI-Syn3.0, containing a genome that had been reduced by almost 550 kb was generated using a bottom-up minimization approach. To generate this strain, Hutchison et al. first compiled a list of essential genes generated largely through prior literature and transposon mutagenesis studies. Genes were then synthesized using PCR and pieced together in yeast to form a contiguous genome. When the synthesized genome was

transplanted into *M. mycoides*, the cells exhibited no viability. Additional innovation in transposon mutagenesis techniques provided the ability to identify an additional class of genes, termed quasi-essential, which are defined as genes which, when deleted, result in an observable fitness defect. The combination of essential genes and quasi-essential genes, following synthesis assembly into a genome in yeast and transplantation into *M. mycoides*, proved to be sufficient to sustain life. The entire process, from inception of the idea to publication, was reported to take about 17 years.

More recently, a similar attempt was made to generate an *E. coli* strain containing a substantially reduced genome. In this study, a putative minimal 1.03 Mb E coli genome was elucidated and then synthesized in yeast[27]. However, transplantation of this genome from yeast into *E. coli* failed to yield a viable cell.

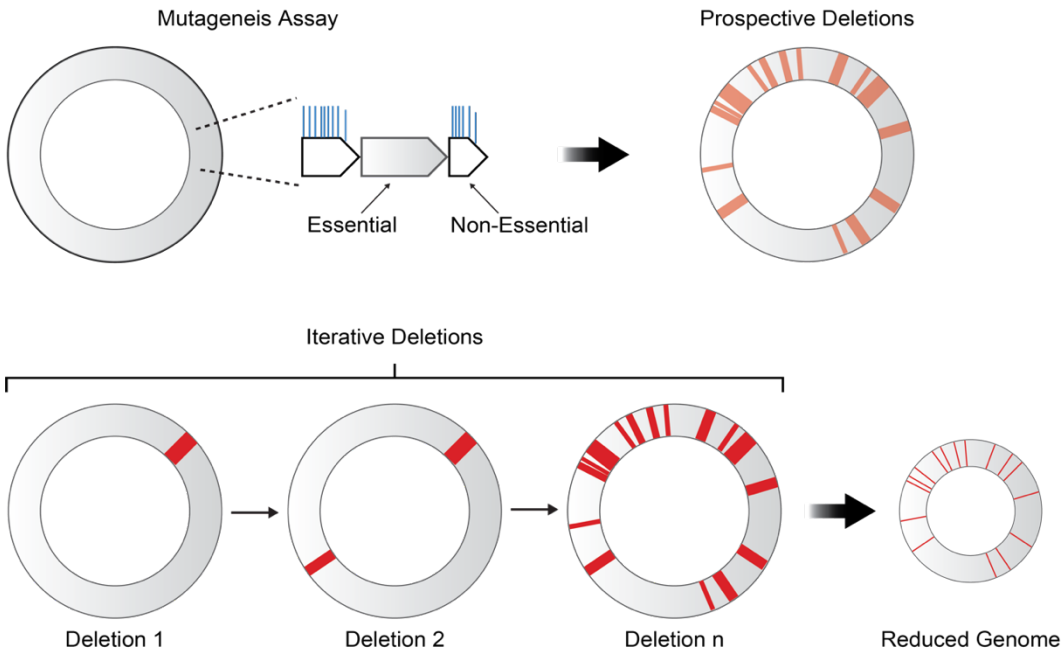

Figure 1.2: Schematic representation of top-down minimization
A typical top-down minimization scheme follows the following steps: (1) Delineate essential from non-essential portions of the genome. (2) Based on information from step 1, determine contiguous regions to be removed and the in which they are to be removed. (3) Iteratively remove non-essential regions until the desired reduced genome has been achieved.

Top-down genome minimization studies comprise a slightly more diverse subset of minimization tactics. One such tactic, directed top-down minimization (**Figure 1.2**), shares some similarities with bottom-up minimization schemes. Directed top-down minimization efforts generally also begin with thorough assessment of essentiality and fitness studies. However, unlike in bottom-up minimization schemes, the studies are used to identify contiguous segments of DNA which can be removed from the genome. In many directed top-down minimization studies, segments of DNA which encode for prophage elements, insertion sequences and other mobile genetic elements will be designated to be removed[28]. In addition, genes defined generally as non-essential will be designated for removal as well. Once a list of pieces designated for removal is constructed, these individual components will be grouped into larger, contiguous segments to be removed. A single contiguous segment of DNA is removed at time. Directed top-down minimization studies often suffer from many of the same shortcomings as bottom-up minimization efforts. Information used in these studies to designate segments for removal is often derived from first order mutagenesis studies or wild-type fitness assessments. As the minimization process proceeds away from the wild-type form of the genome the information derived from those studies becomes less relevant, possibly resulting in incomplete or even incorrect assessments of particular segments of DNA. Incomplete information may prohibit deletion of a target DNA segment, resulting in a forced reevaluation of the segment and moving the study back into the design-build-test cycle.

One of the most renowned directed top-down minimization efforts is that of the multiple deletion series, which generated a widely utilized genome-reduced *E. coli* strain MDS42[29]. MDS42 was generated by making planned and precise deletions, leading to the removal of almost 14.3% of the wild-type genome. Initially, a cross-strain comparison of genomic elements was conducted to identify genomic segments with no overlap between *E. coli* K-12 MG1655 and other *E. coli* strains. Amongst non-overlapping segments, those predominantly made up of insertion

sequences (IS) were deleted first. 42 sequential deletions, carried out using Lambda-Red mediated homologous recombination[30], were made until all IS elements had been removed. After removal of all IS elements, the resulting strain, MDS42, displayed a higher degree of genome stability than wild-type and a similar growth fitness in standard conditions. MDS42 also exhibited a higher electroporation efficiency than wild-type. In succeeding years the multiple deletion series effort advanced even further, removing additional genomic segments to generate MDS69[31], resulting in a strain with 20.3% of the wild-type genome removed across 69 total rounds of genome reduction.

The most genome-reduced *E. coli*, DGF-298 was generated through a series of studies, the first of which resulted in the generation of MGF-01. MGF-01 was generated through the removal of 22% of the wild-type *E. coli* W3110 genome[32], in a manner similar to that as described for the generation of MDS42. Briefly, genomic segments were identified for removal through a genomic comparison between *E. coli* W3110 and *Buchnera sp.*, a symbiont thought to share a common ancestor with *E. coli*[33]. Through this analysis, 83 candidate regions were identified for deletion. An additional 20 regions were identified for deletion by function, including IS elements and toxin-antitoxin pairs. Genomic segments were removed, once again, using Lambda-Red mediated homologous recombination. In total, 53 rounds of deletions were required to produce MGF-01. Following the generation of MGF-01, comparative genomic analyses were conducted to identify genomic segments that had been deleted in other genome reduced strains but not MGF-01. These segments, along with additional IS elements, toxin-antitoxin pairs and prophage elements were identified for deletion. An additional 43 rounds of deletions were conducted to generate DGF-298[34], resulting in a genome-reduced *E. coli* strain with a genome approximately 37.4% smaller than wild-type. DGF-298 exhibits better growth fitness in various media and a higher maximum cell density in rich medium than wild-type.

While directed top-down minimization studies are prevalent in *E. coli*, several studies have focused on generating genome reduced strains of different bacteria. Prior genome annotations were used to identify dispensable components of the *Bacillus subtilis* genome. 94 rounds of deletions were carried out in rich media, resulting in the reduction of the *B. subtilis* genome by 36.5%[35]. In *Streptomyces avermitilis*, a single 1.4 Mb deletion and a series of smaller deletions were made to generate SUKA17, resulting in a genome reduction of 18.5% from wild-type[36]. A combination of homologous recombination and cre/lox recombination was used to generate this strain.

The genome reduction methodologies used in the bottom-up and directed top-down minimization studies, were each deployed in exactly one strain and are highly species specific. While these approaches produced functioning genome-reduced organisms, their reliance on prior knowledge introduced strong biases. For example, gene essentiality is often assessed using transposon mutagenesis in a specific environmental context, ignoring the interconnected and conditional nature of gene function. Translation of the exact methodology used in any of the previously mentioned studies to other bacteria would mean starting over completely from scratch. As these methods rely heavily on prior information, translation of these methods into other, less studied bacteria may be impossible without decades of foundational studies.

The other form of top-down minimization relies on random selection to determine which DNA segments to delete. The utilization of random selection avoids the intrinsic biases associated with a reliance on prior information. As a result, this method expedites the genome reduction process by avoiding the deletion of any necessary components followed by reversion back into the design-build-test cycle. All random top-down minimization studies to date have employed transposition-based systems for target selection. Prior to the development of fully random targeting and deletion methods, pseudo-random deletion methods were employed.

An early study employing this type of method combined random transposition with site-specific recombination to remove genomic segments in *E. coli*[37]. First, transposition was utilized to randomly integrate a cassette containing an antibiotic resistance gene and a loxP sequence. 800 unique insertions were identified, each in an independent cell. This process was repeated in wild-type cells to generate a second set of insertion mutants. Next, specific insertion mutants were selected and combined to form 13 unique strains, each containing two loxP sites and two antibiotic resistance genes. Finally, cre expression was induced and deletions were observed in 6 of the strains. Subsets of deletions were then combined to generate 3 largely overlapping genome-reduced strains with the most reduced having roughly 6% of the wild-type genome removed. More recent random top-down minimization schemes transpose a single nuclease target sequence to a random genomic locus, then activate the requisite nuclease to produce a double stranded break (DSB). In most bacteria, the presence of a DSB will induce the DSB repair system, which will chew back DNA to an undefined extent. The DSB will then need to be repaired by endogenous systems such as homologous recombination[38] or alternative end joining[39] to rescue the cell from death by genomic degradation. Two notable studies employ this technique, RANDEL[40] and TMRD[41].

The RANDEL method for random genome reduction transposes a cassette containing two I-SceI nuclease target sites flanking an antibiotic resistance gene and a counterselection gene to a random genomic locus. Following transposition, expression of the I-SceI nuclease is induced, resulting in the excision of the antibiotic resistance and counterselection genes and the formation of a DSB. As previously described, the DSB repair process will then result in the degradation of endogenous genomic DNA. The RANDEL method was deployed for 5 rounds in *E. coli* MG1655 and 1 round in *E. coli* MDS42, with each round taking 7 days to complete. Across 5 rounds of RANDEL, 60 whole genomes were sequenced and a cumulative total of 12 unique deletions were observed ranging in size from 1.9 to 70.7 kb. Two genome-reduced strains were screened upon completion of 5 rounds

of RANDEL. The two strains were reduced by 2.6% and 2.6% from the wild-type MG1655 genome, with substantially overlapping deletions. Limited evaluation of genome-reduced strains revealed a slight competitive disadvantage in LB medium for the most genome-reduced strain as compared to wt. Interestingly, it was noted that the application of RANDEL resulted in substantial off-target mutations in the genomic background.

The TMRD method employs a similar scheme to RANDEL for random genome reduction. First, a cassette containing a Cas9 target sequence, antibiotic resistance and counterselection genes is transposed to a random genomic locus. Following transposition, Cas9 activity is induced, resulting in the formation of a DSB in the randomly transposed cassette. Like RANDEL, TMRD relies on the endogenous DSB repair process to remove flanking genomic DNA and repair the DSB. TMRD was applied for 5 rounds in *E. coli* MG1655 with each round taking 8 days to complete. Across 5 rounds of TMRD, 20 whole genomes were sequenced and 31 unique deletions were observed, ranging in size from 0.2 to 129.2 kb. Nine genome-reduced strains were screened upon completion of 5 rounds of TMRD. The most genome-reduced strain was reduced by 5.5% from the wild-type MG1655 genome. All nine genome-reduced strains screened following round 5 sustained substantially overlapping deletions. Follow-up evaluations revealed variable fitness for genome-reduced strains as compared to wild-type in both LB and M9 media and improved electroporation efficiencies.

Despite these innovations, both RANDEL and TMRD yielded only limited genome reduction and strain diversity. Each study generated genome-reduced strains with limited diversity after 5 rounds of minimization, removing a maximum of 2.8% and 5.5% of the genome, respectively. The lack of diversity in deletions amongst genome-reduced strains severely limits the potential for follow up characterization and eventual identification of novel biological information including new functions and/or new interactions, two major goals of genome minimization. All deletions

were mediated by endogenous DSB repair pathways, which can vary substantially between bacterial species[38,39,42,43]. Moreover, because each approach was applied only to *E. coli*, their portability remains unproven. Due to these limitations, generating fully reduced genomes using either of these two methods would fail to capture much of the full space of possible fully reduced genomes.

Building upon genome reduction strategies of the past

The genome reduction strategies described thus far compose 3 general categories, bottom-up minimization, directed top-down minimization and random top-down minimization. Both bottom-up minimization and directed top-down minimization depend on prior information to generate genome-reduced strains, resulting in biased outputs and frequent delays due to incompatible deletion attempts. Random minimization has no reliance on prior information and therefore alleviates any biases that are introduced through the use of such information.

Notably, all of the genome reduction strategies described previously have focused predominantly on generating one or a few largely overlapping genome-reduced strains per method and species, which severely limits the biological insights that can be gained. The complex epistatic interactions between genes mean that the removal of one gene can influence the essentiality[44] and expression[45–47] of others. As deletions are combined, gene expression shifts in response, creating feedback loops that influence future deletions. Consequently, the order in which deletions are conducted matters, meaning each deletion order can create unique evolutionary paths toward different viable minimized states, or local minima, potentially illuminating different aspects of genome function and regulation. Each reduced genome generated in previous studies represents only one of potentially many viable reduced configurations contained within the full space of all local minima.

All genome reduction studies described previously were developed for and shown to function in a single species per method. The ultimate goal of genome minimization studies is to identify the foundational building blocks of genomes,

and therefore of life. It is proposed that these foundational pieces could be utilized as a chassis, upon which novel organisms could then be constructed. However, a chassis composed of pieces from a single organism would fail to capture the broad range of foundational pieces that can be elucidated from across the bacterial kingdom. Bottom-up minimization and directed top-down minimization methods have been reported for the generation of genome-reduced strains in a variety of bacteria. However, these studies can take over a decade to complete, are highly species specific and lack the ability to be performed efficiently in parallel. Random top-down genome reduction methods have, thus far, have exhibited no parallelization and have been shown to function only in *E. coli*.

Just as natural selection explores fitness landscapes through variation and selection, synthetic genome reduction should explore multiple viable configurations to fully reveal what is fundamental. To comprehensively understand gene properties such as contextual essentiality and underlying regulatory and functional interactions, it is necessary to generate libraries of genome-reduced strains, each representing a different route through the deletion landscape. Such libraries enable comparative functional omics analyses which can reveal unexpected properties, compensatory interactions, and regulatory rewiring. Furthermore, extending this approach across diverse bacterial species is critical for identifying universally essential elements versus species-specific dependencies. Essentiality is context-dependent, and what is non-essential in one bacterium may be core in another. To identify universal genomic elements that underlie all life and to build generalizable, modular chassis, we must minimize genomes in many different organisms. This requires minimization systems that are efficient, generalizable, and scalable across taxa. Current methods fall short in this regard: they lack the throughput and generalizability to produce genome-reduced strain libraries across multiple species. This represents a major barrier to achieving the full promise of genome minimization.

Chapter 2

WILL THE REAL '**S**TOCHASTIC **L**INEAGE-BASED **I**TERATIVE **M**INIMIZATION' SHADY PLEASE STAND UP

INTRODUCTION

To build upon prior genome minimization schemes, we developed SLIM (Stochastic Lineage-based Iterative Minimization), a modular genome reduction system designed for: unbiased DNA removal, high-throughput parallelization and cross-species compatibility. SLIM utilizes random top-down minimization for genome reduction, thereby avoiding a reliance on prior information and the introduction of biases. Unlike all prior genome reduction strategies, SLIM allows for the reduction of multiple genomes in parallel. Because of this, we are able to generate diverse libraries of genome-reduced strains. The generation of libraries of genome-reduced strains allows for comprehensive multi-omics analyses, comparing expression patterns of the remaining genes amongst genome-reduced bacteria. The species versatility of the components used in SLIM for DNA targeting and removal allow for a seamless transition from reducing the genome of *E. coli*, the host used for the development of SLIM, to reducing the genomes of other phylogenetically distinct bacteria.

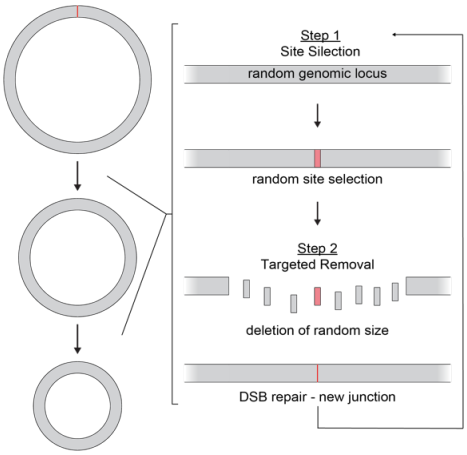

Figure 2.1: SLIM-mediated genome reduction overview
Sequential iterations of SLIM are used to reduce genomes. Each iteration of SLIM consists of two steps: In step 1 of SLIM, a genomic locus is randomly designated for removal. In step 2 of SLIM, an undefined portion of the selected locus is removed, creating a DSB. The next round of SLIM can begin following DSB repair.

SLIM proceeds in two sequential steps, iteratively reducing the genome (**Figure 2.1**). First, a target cassette composed of exogenous DNA is randomly inserted into the genome. Next, the cassette and flanking endogenous DNA are degraded by a guided nuclease system. The amount of genomic DNA removed is unspecified; it varies from cell-to-cell and by genomic location. This process is repeated iteratively, each time shrinking the genome until all non-essential DNA has been removed, yielding a fully minimized genome.

Step 1 of SLIM is mediated by the Tn5 transposition system. This system comprises a Tn5 transposase and a transposon cargo composed of two transposase recognition sites, Tn5 mosaic ends (Tn5ME), flanking the transposon cargo[48]. The Tn5 transposition process is completed in 6 steps. First, a tn5 transposase monomer binds each Tn5ME on the transposon cargo. Next, the Tn5 monomers complex and mediate a nucleolytic attack on the DNA molecule containing the transposon cargo, excising the cargo from its original position. Since Tn5 is a "cut-and-paste" transposition system, a DSB is formed in the original host molecule at the excision site. Following excision, the transposase/transposon cargo complex binds to a random target sequence and initiates a nucleophilic attack at the target site forcing a strand exchange between the transposon cargo and the target DNA molecule. The transposition process is completed upon release of the cargo. The Tn5 transposase used in this study is a hyperactive variant, designed to be more efficient by the induction of mutations which enhance Tn5 transposase binding efficacy and block inhibition activity[48]. Because of its seemingly unbiased nature, the Tn5 transposition system is widely used for both *in vitro* and *in vivo* applications. Tn5 mutagenesis studies have shown an insertion site selection diversity approaching 10^6 unique sites, resulting in an average of 1 insertion per every 5.14 bp[49]. In addition, the Tn5 system is commonly used for size specific DNA fragmentation for downstream DNA sequencing applications[50].

Step 2 of SLIM is mediated by a programmable Type 1-C CRISPR-Cas cascade system anchored by Cas3, a 3′–5′ single-strand DNA helicase-nuclease. The Cas3-cascade is composed of 3 additional proteins, Cas5, Cas7 and Cas8. These 3 proteins form the crRNA-guided surveillance complex, which identifies and binds specifically to the target sequence. Once bound to the target sequence, the complex can recruit Cas3, which binds to the strand opposite that of the complex. The binding of Cas3 initiates a processive 3′→5′ degradation of surrounding genomic DNA for an unspecified distance before switching strands and excising DNA in the opposing direction. The deletion size and boundaries vary, adding to the stochastic nature of the method. Unlike other random genome-reduction methods that utilize enzymatic systems which simply induce DSB and rely on host repair pathways to remove endogenous genomic DNA, Cas3 performs continuous processive degradation. This activity is intrinsic to the system and does not depend on the host's repair machinery. As a result, Cas3 is less sensitive to species-specific differences in repair mechanisms, enabling broader applicability. Cas3 has demonstrated efficacy in multiple bacterial species, enabling large (100+ kb) deletions in *E. coli*, *Pseudomonas aeruginosa*, *Pseudomonas syringae*, and *Klebsiella pneumoniae*[51]. In SLIM, the Cas3-cascade is driven by a rhamnose inducible expression system, encoded on pCas3, alongside a modified version of the requisite crRNA.

SLIM transforms genome minimization from a single-strain pursuit to a systems-level exploration of the genomic design space. We applied SLIM to *E. coli* for 21 iterative rounds of genome reduction, generating a library of 10 genome-reduced strains. We then performed detailed phenotypic and molecular characterization of this library, revealing new insights into deletion permissibility, the effects of deletions on expression dynamics, and context dependent phenotypic behaviors driven by expression dynamics. To demonstrate modularity, we deployed SLIM in two additional species: *Shigella flexneri* and *Pseudomonas putida*. In both cases, SLIM successfully generated diverse libraries of genome-reduced strains,

underscoring its generality and potential as a broadly applicable tool for bacterial genome minimization. In this study we demonstrate that SLIM provides a scalable, unbiased, and generalizable platform for genome minimization. By enabling parallel, iterative genome reduction across species, SLIM unlocks the full landscape of minimized genomes and moves us closer to identify components for universal genetic chassis.

RESULTS

Genome Reduction: Description and Analysis of Steps

The genome minimization process employed in SLIM proceeds in two core steps: (1) Delivery and Transposition, and (2) DNA Removal and Repair.

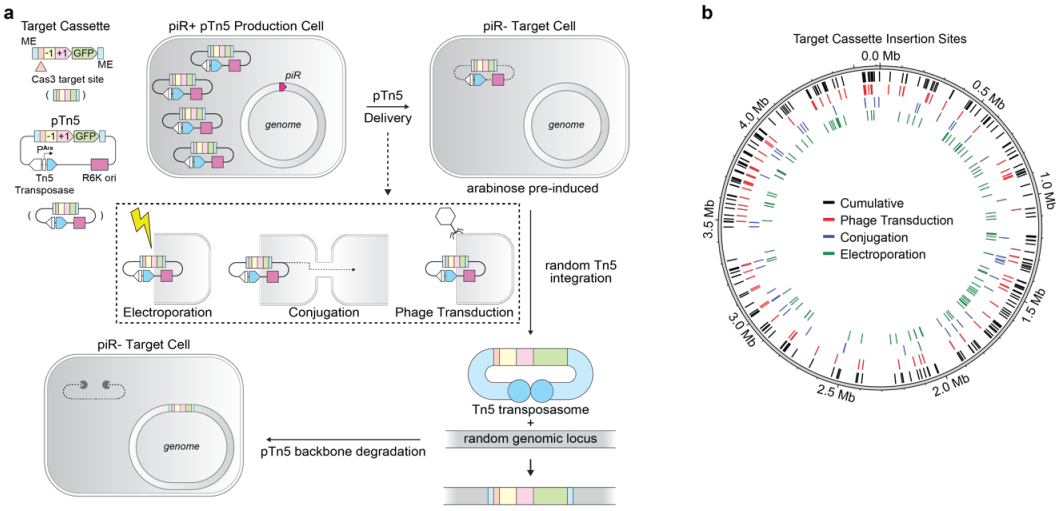

Figure 2.2: SLIM step 1, delivery and transposition
(a) Schematic depicting step 1 of SLIM, pTn5 delivery and random target cassette integration. The three modes of pTn5 delivery utilized throughout the development and deployment of SLIM were electroporation, conjugation and phage transduction (dotted rectangle). (b) Results of a modified TnSeq assay used to identify unique landing sites, plotted by genomic location, of the target cassette following delivery and transposition, stratified by delivery mode (phage transduction-red, conjugation-blue and electroporation-green) and the combined locations identified for all delivery modes (cumulative-black).

Step 1: Delivery and Transposition

In the first step, the plasmid pTn5 is introduced into *E. coli* target cells that have been pre-treated with arabinose. Delivery is achieved via one of three methods: electroporation, conjugation, or P1 phage transduction (**Figure 2.2a**). The pTn5 plasmid encodes an arabinose-inducible, hyperactive Tn5 transposase, which is expressed immediately upon plasmid entry due to arabinose pre-treatment. This transposase mediates the excision and random genomic integration of a synthetic target cassette encoded by pTn5.

The target cassette includes a 34-bp target sequence derived from the *E. coli* gene pdeL, a hygromycin resistance gene (hygR), a mutant allele of pheS (pheS T251A_A294G, hereafter pheS^{mut}) for counterselection, and the GreenLantern variant of GFP[52]. pTn5 relies on the pi-replicator (piR)/R6Kɣ system for replication[53]; chosen to prevent plasmid propagation in target cells due to the absence of the piR protein, thereby enriching for cells that have stably integrated the target cassette into their genome.

The transposition and delivery processes are completed simultaneously in liquid culture. Cultures are then plated on LB-agar containing hygromycin to select for successful cassette integration. To evaluate delivery efficiency and insertion site diversity, we tested each delivery method in parallel. Approximately 200 colonies were collected for each method, pooled, and analyzed using a modified TnSeq approach[54]. Each method—phage transduction, conjugation, and electroporation—successfully facilitated broad distribution of cassette insertions across the genome (**Figure 2.2b**), yielding 86, 47, and 96 unique insertion sites, respectively.

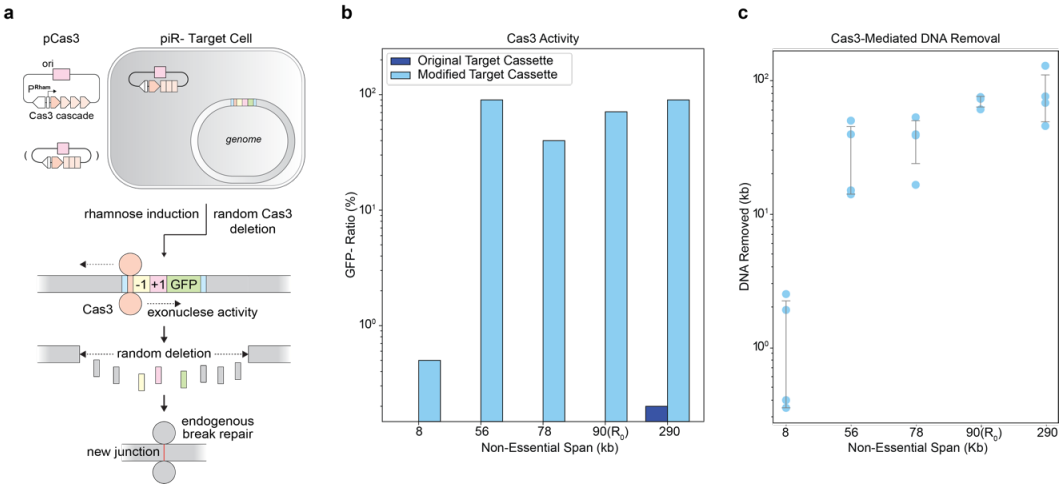

Figure 2.3: SLIM step 2, DNA removal and DSB repair
(a) Schematic depicting step 2 of SLIM, Cas3-mediated DNA removal and endogenous DSB repair. (b) Categorical bar plot depicting Cas3 activity following integration of either the original target cassette (dark blue, not encoding pheS^{mut}) or modified target cassette (light blue, encoding pheS^{mut}) at select genomic locations, identified by the size of the non-essential span (distance between the nearest essential genes flanking the target cassette insertion). 90(R_0) indicates the non-essential span and the set from which the progenitor strain was derived. (c) Categorical scatter plot depicting the

amount of endogenous DNA removed from the genome of cells exhibiting Cas3 activity following modified target cassette integration, stratified by non-essential span.

Step 2: DNA Removal and Repair

Following target cassette integration, DNA is removed by the action of the CRISPR-Cas3 system encoded on the pCas3 plasmid. This system uses a multi-protein Cascade complex and the Cas3 helicase-nuclease to identify and degrade DNA containing the programmed target sequence. Expression of the CRISPR-Cas machinery is induced with rhamnose. Upon activation, Cas3 processively degrades the target cassette and flanking genomic DNA (**Figure 2.3a**). Resulting double-strand breaks are typically repaired by the cell's dominant endogenous DNA repair machinery; recA-mediated homologous recombination for *E. coli*[51].

To evaluate the efficacy and dynamics of Cas3-mediated DNA removal in conjunction with target cassette landing site, we inserted target cassettes into five distinct non-essential genomic regions of varying size (8 kb, 56 kb, 78 kb, 90 kb, and 290 kb), equidistant from essential genes, using Lambda-Red recombination[30]. These target sites were chosen to span a range of conditions, from a small non-essential span (8 kb) expected to yield short segments of removed DNA and low DNA removal efficiency to a large non-essential span (290 kb) expected to yield large segments of removed DNA and high DNA removal efficiency.

Initial tests using a simplified version of the cassette (target sequence + hygR + GFP) yielded few to no GFP- colonies after Cas3 induction, suggesting limited to no Cas3 activity and resulting target cassette degradation (**Figure 2.3b, dark blue**). To enhance Cas3-mediated target cassette degradation, we incorporated $pheS^{mut}$ into the target cassette, forming the modified target cassette. In the presence of 4-chlorophenylalanine, $pheS^{mut}$-expressing cells incorporate this toxic analog into proteins in place of phenylalanine, causing rampant protein misfolding and eventual cell death[55]. The addition of this counterselection system dramatically increased the efficiency of target cassette degradation across all five loci (**Figure**

2.3b, light blue), with two of the five sites yielding degradation efficiencies above 90%.

To quantify the extent of DNA removal, we selected four colonies per site following Cas3 induction, isolated genomic DNA, and performed whole-genome sequencing. Deletion sizes correlated with the sizes of non-essential spans, ranging from a 0.35 kb deletion in the 8 kb region to a 129 kb deletion in the 290 kb region (**Figure 2.3c**). One strain generated from targeting the 90 kb span, which exhibited approximately 60 kb of DNA loss, was chosen as the progenitor (R_0) to which SLIM was applied to generate the library of genome-reduced strains (**Figure 2.12d**). The removal of DNA to generate R_0 is hereby referred to as Round 0.

Application of the full SLIM cycle to R_0 yielded ten genome-reduced strains, with eight of the strains bearing a unique deletion. The genomic locations of these deletions collectively span over 3 Mb of the *E. coli* genome, ranging from 0.756 Mb to 3.879 Mb. Interestingly, deletions in colonies 1, 2, and 5 overlap within the same general genomic region, yet the extent of deleted DNA differs substantially, 36 kb, 18.6 kb, and 40.8 kb, respectively. This result demonstrates that distinct Cas3 degradation events can converge on overlapping regions while producing unique deletion outcomes, further underscoring the stochastic and flexible nature of the SLIM system.

The application of SLIM in *E. coli* produces a library of genome-reduced strains

Previous genome minimization efforts have typically aimed to isolate a single "optimal" strain with the smallest possible genome, often by applying competition-based selection strategies. While effective for achieving deep reduction, such approaches inherently limit the diversity of genomic outcomes. Due to the complex network of interactions within bacterial genomes, there are many distinct paths that can lead to viable, fully reduced genomes, each representing a local minimum. However, as deletions accumulate through the genome reduction process and the

diversity of deletions is reduced through competitive culture conditions, the number of possible paths is reduced and the space of possible fully reduced genome configurations becomes increasingly constrained (**Figure 2.4a**).

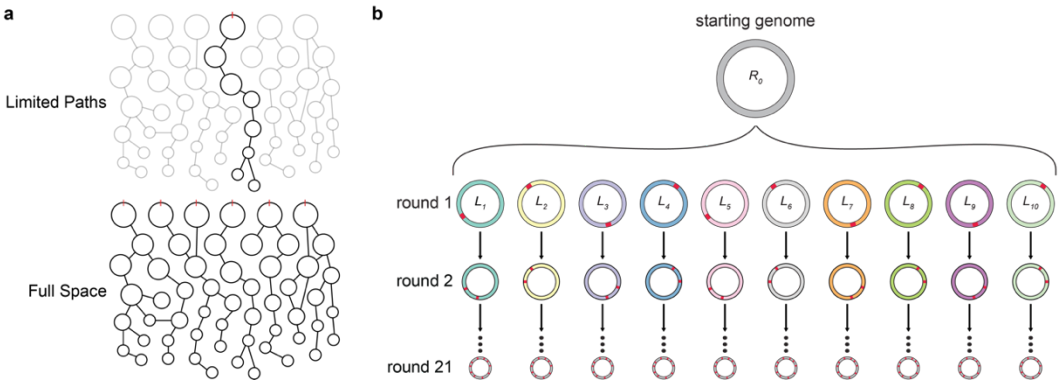

Figure 2.4: SLIM library parallelization schematic
(a) Illustrative representation of the full space of paths from a wild-type genome to a fully minimized genome, following iterative deletions. Propagation of a single deletion (indicated in red) through iterations limits the number of possible paths and resulting variations of minimal gnomes ('Limited Paths'). The propagation of multiple, independent deletions in parallel preserves the possible paths, expanding the output of fully reduced genomes ('Full Space'). (b) Illustrative representation of the library scheme used in the application of SLIM for 21 rounds in *E. coli*. Ten lineages derived from R_0 were generated during Round 1. DNA was removed iteratively and independently in parallel for each lineage in subsequent rounds.

To address this limitation, we implemented a library-based strategy in which multiple genome-reduced strains were generated independently and in parallel. This design preserves path diversity and enables the exploration of a broader landscape of deletions and genetic interactions, thereby enhancing our ability to infer underlying relationships between genes and pathways.

We applied four variations of the SLIM workflow across 21 total rounds of genome reduction, ultimately generating a library of 10 unique genome-reduced *E. coli* strains (**Figure 2.4b**). While Tn5 transposition and Cas3-mediated DNA removal parameters remained largely consistent, we varied pTn5 delivery methods and timing across rounds (**Figure 2.12a-c**). Rounds 1–3 used electroporation for delivery of pTn5. When utilizing electroporation as the pTn5 delivery method, each round of SLIM took 5 days to complete. Rounds 4–16 used conjugation (4 days/round), rounds 17–20 used P1 phage transduction (4 days/round), and round

21 also used P1 phage transduction, but featured optimized timings that reduced the workflow to 3 days.

For Round 1, pTn5 was delivered to a monoculture of the progenitor strain (R_0). Following delivery and transposition, 20 independent colonies were selected and pooled into 10 pairs, each representing an independent lineage. Cas3-mediated DNA removal was then performed independently for each lineage. Across subsequent rounds, two colonies per lineage were typically selected for sequencing and to be pooled for use as a seed for the next round. All 10 strains were sequenced after rounds 1, 2, 3, and 5; nine out of ten were further sequenced after rounds 10 or 12, 15, 20, and 21. Strain 5 was excluded from later rounds due to a prohibitive growth deficiency observed during rounds 6 and 7.

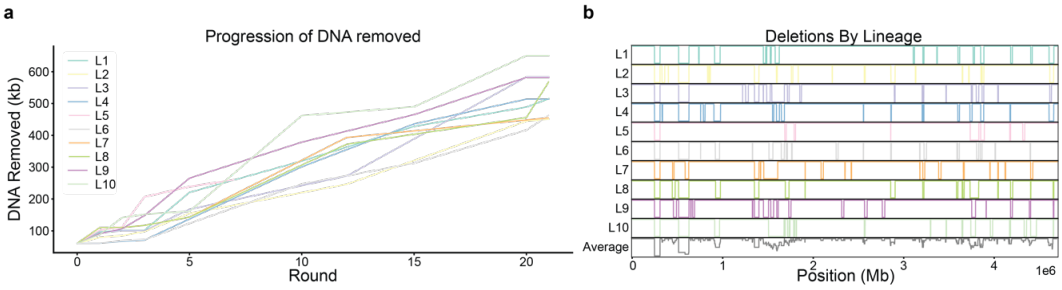

Figure 2.5: Examination of DNA removed though 21 rounds of SLIM
(a) Line plot depicting cumulative DNA removed, per lineage, accumulated by round through 21 rounds of SLIM in *E. coli*. (b) Coverage plot representing all observed deletions accumulated for each lineage following 21 rounds of SLIM. Deleted portions are given a value of 0 and remaining portions are given a value of 1. The position and magnitude of every observed deletion is represented, per lineage. The 'Average' row represents the average DNA removed on a per-nucleotide basis across all ten lineages.

Across 21 rounds, we observed 16 to 19 deletions per strain, with deletion sizes ranging from 75 bp to 170 kb (**Appendix A**). The cumulative DNA removed per strain ranged from 434 kb to 649 kb, representing 9.25% to 13.9% of the parental *E. coli* DH10B genome (**Figure 2.5a**). Deleted segments spanned the entire genome, and, aside from the original deletion in R_0, there were no shared deletion events across all ten strains (**Figure 2.5b**). Nine of ten strains had lost at least part of one of the two 113 kb tandem repeat regions in DH10B (spanning approximately

0.5–0.75 Mb), a known genomic feature[56]. Additional regions with high or low deletion frequency were observed as evidenced in **Figure 2.5b** ('Average' row).

Understanding deletion patterns amongst genome-reduced *E. coli*

To better understand the factors influencing deletion permissibility, we analyzed the final set of deletions generated across all genome-reduced strains. We hypothesized that both gene function and genomic location contribute to a gene's likelihood of being deleted and examined each characteristic in isolation.

To assess the role of gene function, we classified all genes into 25 Clusters of Orthologous Genes (COG) functional categories[57]. Qualitative inspection revealed that certain categories, such as 'Inorganic ion transport and metabolism,' 'Transcription,' and 'Secondary metabolite biosynthesis,' showed high levels of deletion permissibility. In contrast, categories such as 'Translation, ribosomal structure, and biogenesis' and 'DNA replication, recombination, and repair' exhibited markedly lower deletion frequencies (**Figure 2.6a**). This trend was further supported by analyzing the median deletion percentages per COG group (**Figure 2.6b**). While some groups clearly resisted deletion or were highly permissive, most fell within a moderate range, suggesting that gene function plays a nuanced but variable role in deletion permissibility.

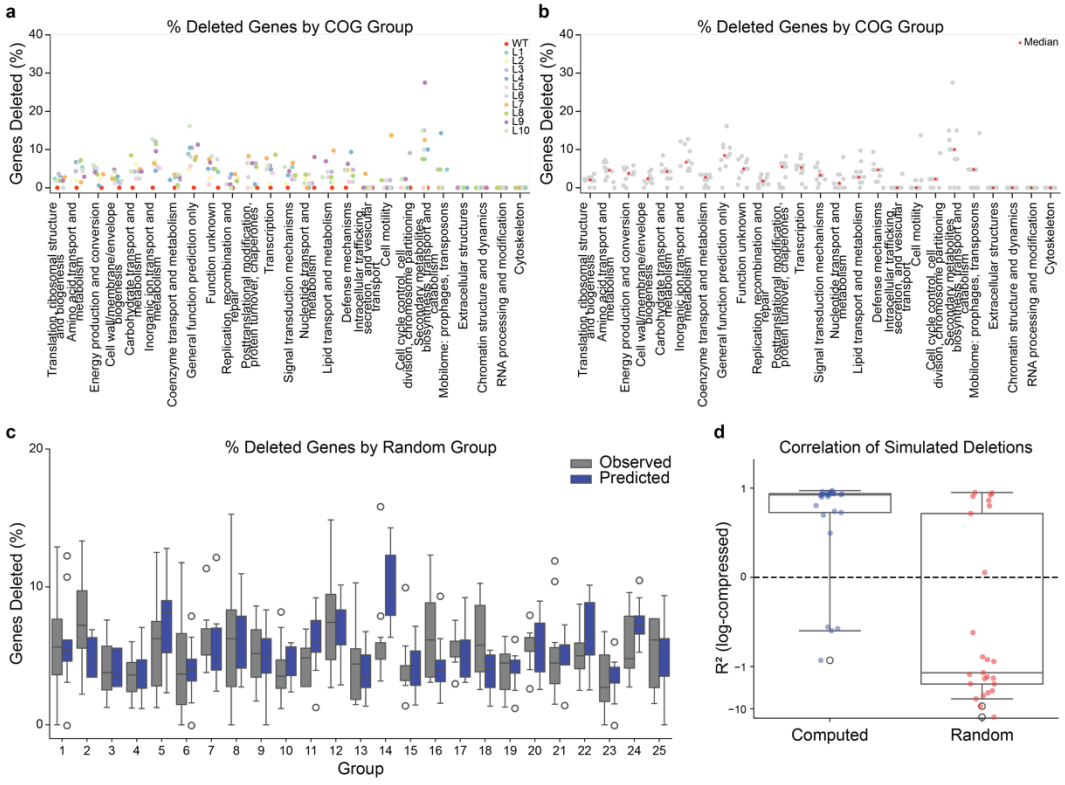

Figure 2.6: Deletion permissibility analyses
(a) Categorical scatter plot stratified by 25 COG groups, depicting the percentage of genes deleted for each genome-reduced *E. coli* strain and WT, per group, colored by strain. (b) Categorical scatter plot stratified by 25 COG groups, depicting the percentage of genes deleted for each genome-reduced *E. coli* strain and WT (gray), per group, with the median percentage of genes deleted per group (red). (c) Categorical box plot comparing the percentage of observed deletions (gray) amongst all 10 genome-reduced *E. coli* for each of the 25 randomly generated groups to the percentage of predicted deletions (blue) compiled from 10 simulations. (d) Categorical boxplot depicting the computed correlations of predicted deletions generated through simulations to observed deletions. Correlations are shown separately for deletions predictions generated based on computed deletion probabilities (blue) and randomly assigned deletion probabilities (red).

We next examined the role of genomic location on deletion permissibility, independent of gene function. For this assessment, we simulated deletions based on assigned deletion probabilities using only genomic location and previously observed deletions as prior information. To do this, we divided the genome into 500 bins based on position and calculated the average amount of DNA removed per nucleotide within each bin using all deletions observed across the 21 rounds of SLIM (**'Average' row, Figure 2.5b**). Each gene was assigned a deletion probability based on the computed average for its bin and randomly assigned to one of 25 "random groups." We then conducted 10 *in silico* pan-genome deletion

simulations, predicting the proportion of deleted genes per group, and compared predictions to actual observations (**Figure 2.6c**). For variations of this assessment we altered the number of bins used to partition the genome (**Figure 2.17**).

To quantify model performance, we computed the coefficient of determination (R^2) between predicted deletions, generated using computed deletion probabilities, and observed deletions across groups (**blue dots, Figure 2.6d**). R^2 values varied but were generally positive, indicating the meaningful predictive power of genomic location for determination of deletion permissibility. For comparison, we randomized the deletion probabilities across bins while maintaining group assignments and performed 10 additional simulations. These randomized controls yielded substantially lower or negative R^2 values (**red dots, Figure 2.6d**). Overall, average R^2 values were 0.47 for computed deletion probability-based predictions vs. –2.2 for randomized deletion probability-based predictions, suggesting that our model utilizing genomic location-based deletion probabilities is a reliable predictor of deletion permissibility.

Finally, we examined whether observed deletions included genes previously identified as essential. Using three independent essential gene datasets[49,58,59], we compiled a list of consensus essential genes (found essential in all three studies) and additional genes found essential in one or two studies (**Appendix B**). Across 21 rounds of SLIM in 10 genome-reduced strains, no consensus essential genes were deleted and each strain deleted an average of only 3.9 non-consensus essential genes. An additional, cursory analysis of toxin/antitoxin (TA) pairs was conducted. Within a 25kb segment of the genome (1.671-1.696 Mb), a cluster of deletions from L3,4,5,6,8 and 10 occupy over 23.2kb. The remaining 1.8kb is composed entirely of the hipA/B TA system. Taken together, these findings further support our conclusion that deletion permissibility is governed by a combination of factors, most notably, essentiality status, genomic location, gene function and TA pairs. Together, these results demonstrate that a multi-lineage, library-based approach

enables not only the generation of diverse genome-reduced strains but also a richer understanding of the forces that govern genome architecture and its constraints.

The effects of deletions on remaining genes are substantial and wide-ranging

Due to the complex interdependencies among genes in the *E. coli* genome, deletion of a single gene can have extensive downstream effects on the remaining genomic components. While the deletion of an individual endogenous gene ('*gene i*') predictably eliminates RNA and protein expression of *i*, the broader cellular consequences are often nontrivial (**Figure 2.7a**). The loss of both RNA and protein *i* can alter the expression levels of many related genes ($j_1, j_2, ..., j_n$), triggering broad transcriptional and translational shifts. Our multi-omics analyses confirm this: across all genome-reduced strains, RNA expression changes were observed in 33.1%–63.1% of genes and protein expression changes in 17.2%–41.1%, despite only 5.1%–16.3% of genes being deleted (**Figure 2.7b**).

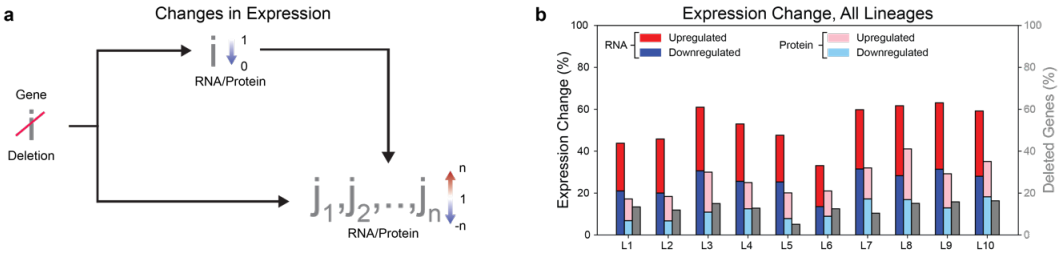

Figure 2.7: Effects of deletions on the expression of remaining genes
(a) Schematic representation of the effects of a single deletion. The deletion of '*gene i*' results in the complete loss of expression of '*gene i*' RNA and protein products and a change in RNA and protein expression of '*genes* $j_{1\text{-}n}$.' (b) Bar plot depicting the percentage of genes exhibiting differential RNA (blue to red) or protein (light blue to pink) expression, per lineage, as compared to the percentage of genes deleted (gray).

We performed both RNA and protein sequencing on all ten genome-reduced strains after 21 rounds of SLIM, as well as on a wild-type control strain (WT, *E. coli* DH10B + pCas3). For each strain, three biological replicates were grown to saturation and harvested. RNA-seq reads were aligned to a DH10B transcriptome (NCBI RefSeq, GCF_000019425.1) and quantified using Kallisto[60]. Differential expression (DE) analysis compared each strain to WT and classified transcripts as

upregulated ($log_2FC > 1.0$, $p < 0.05$), downregulated ($log_2FC < –1.0$, $p < 0.05$), or unchanged. Protein abundance was similarly analyzed. Peptide reads were mapped to DH10B protein-coding sequences (NCBI RefSeq, GCF_000019425.1), quantified, and assessed for DE using the same statistical thresholds. Volcano plots summarizing DE results, including the top five up- and downregulated transcripts/proteins for each strain, are provided in **Figures 2.13 and 2.14**.

To investigate functional consequences of gene expression shifts, we calculated the proportion of total RNA and protein counts attributable to each of the 25 COG categories for every strain. Transcriptome-level accounting revealed that genome-reduced strains allocate a larger fraction of RNA expression to genes involved in 'Chromatin structure and dynamics' compared to WT (**Figure 2.8a**), suggesting an enhanced emphasis on DNA protection, e.g., via genes like hns[61], dps[62], and hupA[63]. This analysis also revealed that WT devotes over 40% of its transcriptomic output to 'Translation, ribosomal structure, and biogenesis,' a category consistently downregulated, in some cases substantially, amongst genome-reduced strains. This trend persists at the proteomic level, albeit less dramatically (**Figure 2.8b**). Variations among genome-reduced strains were also seen in 'Intracellular trafficking,' 'Cell wall/membrane/envelope biogenesis,' and other categories. By contrast, all strains, including WT, show minimal and stable expression in categories like 'Defense mechanisms,' 'RNA processing and modification,' and 'Cell motility.'

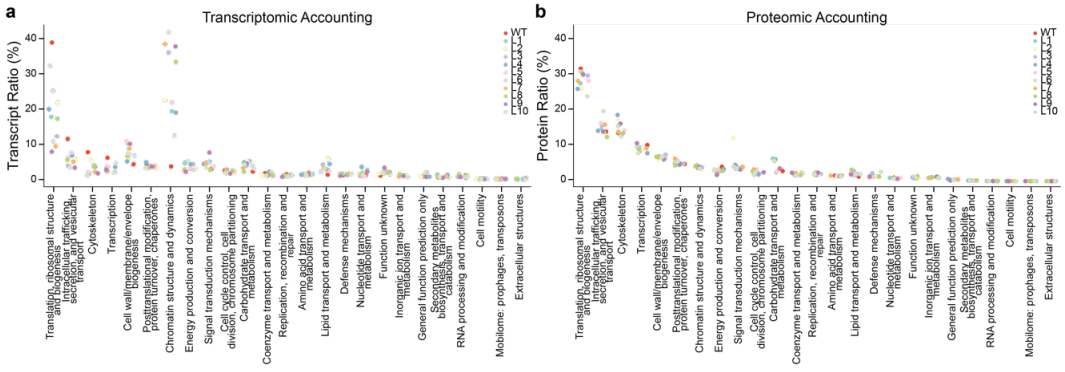

Figure 2.8: Transcriptomic and proteomic accounting analysis across functional groups

(a) Percentage of transcripts (out of total transcripts per strain) counted for genes associated with each of 25 COG groups for each individual genome-reduced strain and WT (red). (b) Percentage of protein products (out of total protein products per strain) counted for genes associated with each of 25 COG groups for each individual genome-reduced strain and WT (red).

To explore changes in specific pathways, we compared RNA and protein DE values for each strain versus WT, focusing on processes such as amino acid transport and metabolism, other metabolism pathways and responses to a variety of stressors. Pathway-level expression shifts were visualized as temperatures derived from a score representing cumulative $\log_2$FC values (**Figure 2.9a**). Pathway enrichment was determined using a Fisher's exact test and indicated for pathways with a p-value < 0.05 (marked with red outlines). Parallel analyses were performed with and without penalties for deleted genes: in the "penalized" case, deleted genes were assigned the lowest observed $\log_2$FC (–10), while in the "non-penalized" case, they were excluded from the analysis. Amino acid transport and metabolism pathways were further stratified by biosynthetic complexity into easy to synthesize (1-2 steps) and difficult to synthesize (greater than 4 steps) groups[64].

Amongst amino acid transport and metabolism pathways, glutamate transport and threonine metabolism pathways were downregulated in all strains except strain 5 according to both RNA and protein DE analyses, regardless of deletion penalty assessment. In the case of glutamate transport, mediated by gltIJKLS[65–67], this trend is likely explained by the genes locations in the tandem repeat region of the DH10B genome, one copy of which is deleted in all strains except strain 5. Genes regulating threonine metabolism, thrABCL and tdcABCDEFG[65–67], are not located in this repeat region. Yet, the patterns of expression shifts in this pathway suggest the possible presence of an uncharacterized regulator within the deleted tandem repeat segment.

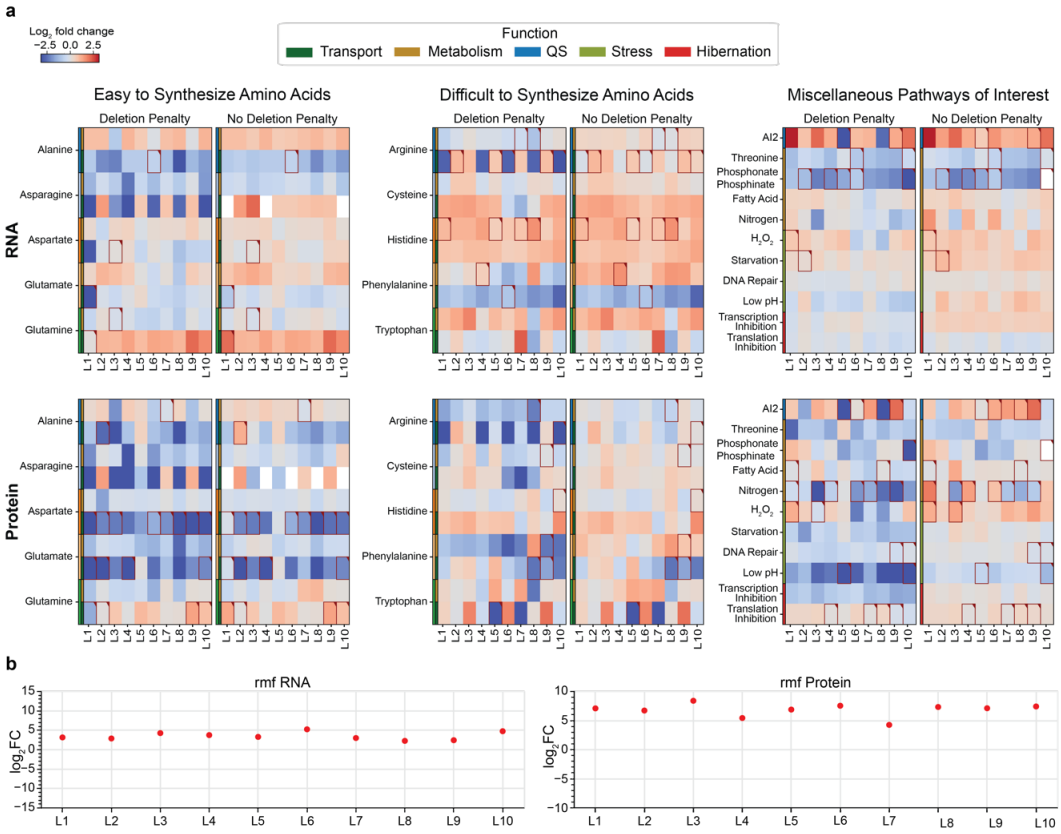

Figure 2.9: Differential expression analysis across individual pathways and genes
(a) Heatmaps depicting the extent of differential expression of select pathways, as total pathway expression, for each genome-reduced strain as compared to WT. The temperatures indicate the extent of upregulation (red) or downregulation (blue) for each pathway as compared to WT, according to either RNA (top row) or protein (bottom row) sequencing output. Heatmaps are broken up into three categories: easy to synthesize amino acids, difficult to synthesize amino acids and miscellaneous pathways of interest. Each category is further divided into two subcategories for which parallel analyses were conducted: 'deletion penalty' and 'no deletion penalty.' 'Deletion penalty' computations apply the lowest observed $\log_2$ fold change (FC) value to deleted genes in total pathway expression computations. 'No deletion penalty' computations simply exclude deleted genes from total pathway expression computations. Pathway identifiers are indicated on heatmaps, and functional information is indicated by colored rectangles following pathway identifiers. (b) Log_2FC for rmf transcripts (right) and proteins (left), as compared to WT, for each genome-reduced strain. All $\log_2$FC values have been normalized to fit between -3 and 3.

Other pathways of interest also were also differentially expressed. For example, DE analysis revealed consistent upregulation of the hydrogen peroxide (H_2O_2) stress response and starvation stress response across most genome-reduced strains, suggesting an emphasis on adaptation to metabolic and oxidative stress conditions.

Comparison of penalized versus non-penalized pathway scores yielded insights into deletion-driven regulation. In some cases, such as cysteine transport in lineages 6–8 and auto-inducer II (AI2)-mediated quorum sensing in lineages 5 (just RNA) and 8 (both RNA and protein), penalized scoring suggested downregulation due to gene deletions, while non-penalized analyses revealed upregulation among the remaining, intact genes. These findings highlight a key distinction: while whole pathways may be affected by deletion events, surviving genes may undergo compensatory regulation. Conversely, some pathways exhibit consistent regulation across all strains, regardless of deletion: glutamine transport and histidine metabolism are generally upregulated, while alanine and phenylalanine transport and phosphonate/phosphinate metabolism are consistently downregulated.

Finally, one especially striking observation emerged from DE analyses: all genome-reduced strains exhibited substantial upregulation of rmf, encoding the ribosome modulation factor (**Figure 2.9b**). This protein promotes ribosome hibernation and stabilization[68], suggesting a general shift in translational strategy and resource conservation in genome-reduced strains.

Examining the interplay between environmental context and differential expression

In the previous section, we examined how genome reduction alters gene expression at the transcript and protein levels. Here, we extend our analysis to explore how these expression changes affect overall cellular fitness. *E. coli*, like many organisms, has evolved a complex genome that supports survival across a wide range of environments, with genes that mediate responses to changes in nutrient availability, osmolarity, pH, temperature, oxygen levels, and other environmental conditions[69–71]. As a result, genome reduction outcomes are heavily influenced by the environment in which reduction occurs[72]. All data presented thus far derive from SLIM-mediated genome reduction and follow-up analyses conducted under standard culture conditions, LB medium at 37°C with appropriate antibiotics.

Utilizing SLIM for genome reduction in different environments would likely yield divergent genome-reduced strains.

As previously discussed, gene deletions in genome-reduced strains lead to substantial changes in RNA and protein expression profiles of non-deleted genes. These expression changes, in turn, alter cellular fitness. This relationship can become especially pronounced as a result of environmental perturbations. Changes to the external environment (e.g., nutrient supplementation) can dramatically affect growth, often by exacerbating or rescuing fitness-altering expression changes (**Figure 2.10a**). In the following experiments, we evaluated how such perturbations interact with expression changes to shape the growth behavior of genome-reduced *E. coli.* We also identified candidate expression changes that likely mediate observed phenotypes.

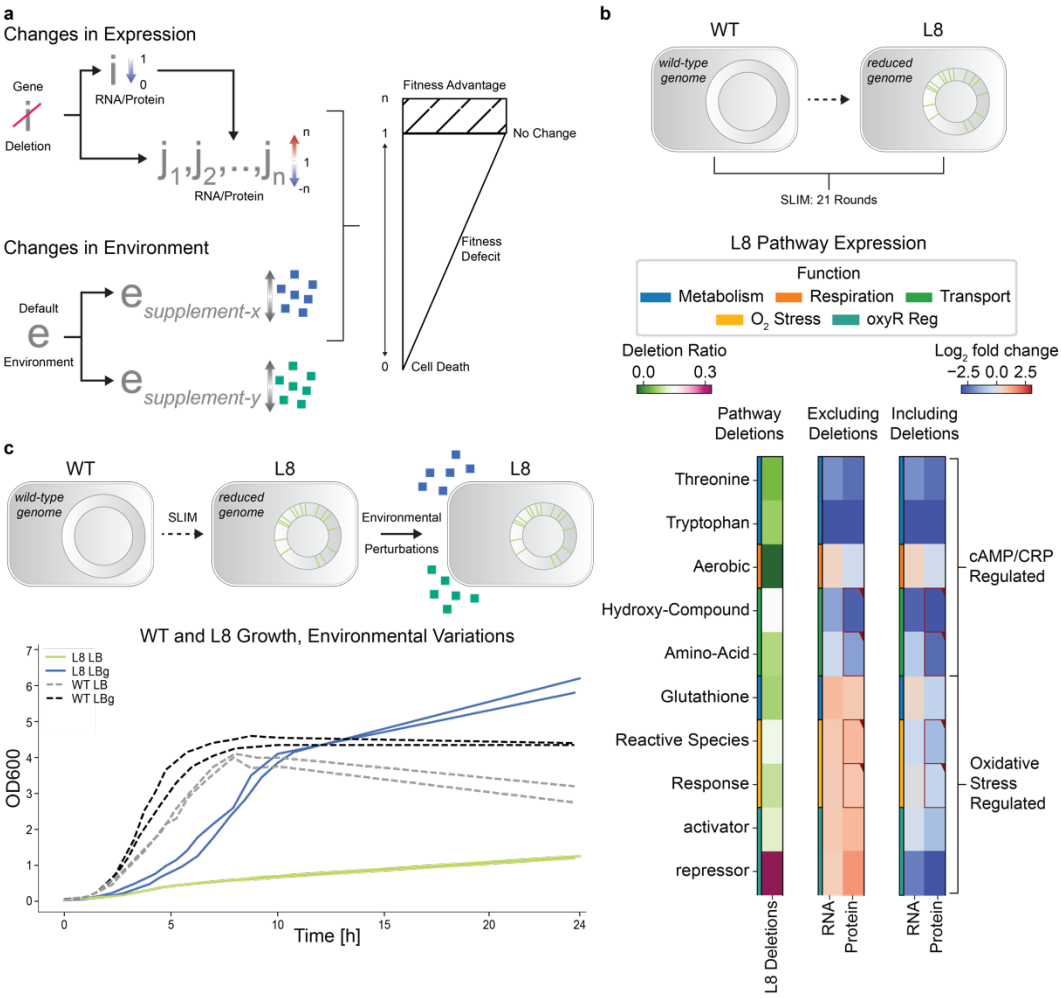

Figure 2.10: Analysis of fitness changes following environmental perturbations due to wholesale expression changes

(a) (top left) Changes in expression derived from deletions, as described in figure 2.6a. (bottom left) Schematic depiction of possible modes of environmental alteration. The external environment can be altered through the addition or removal of elements (blue and green squares). (right) Changes in expression combined with changes in environment can yield dramatic changes in overall cellular fitness, ranging from severe fitness defects and cell death to fitness advantages (dashed rectangle). (b) (top) Schematic depicting the start and end points of 21 rounds of SLIM applied to WT to generate the genome-reduced strain, L8, including markings indicating each new genomic junction generated in the genome of L8 (green) following the removal of genomic DNA. (bottom) Heatmaps depicting deletions sustained and differential expression of select pathways measure based on RNA and protein sequencing output for L8 as compared to WT. The ratio of genes deleted from each pathway to the total number of genes in each pathway, for L8, is presented in a green (0) to magenta (0.3) temperature gradient. Differential expression heatmaps are presented using a temperature gradient and segregated into subcategories as described in figure 2.8a, with separate heatmaps for 'deletion penalty' and 'no deletion penalty' computations. Each set of heatmaps has been further segregated into two functional categories: 'cAMP/CRP Regulated' and 'Oxidative Stress Regulated.' (c) (top) Schematic depicting the start and end points of 21 rounds of SLIM applied to WT to generate the genome-reduced strain, L8 as shown in (b). Following genome reduction, the external environment of L8 was perturbed through a nutrient supplementation and fitness was evaluated as compared to WT. (bottom) Growth curves of both L8 (solid) and WT (gray-scale,

dashed) in standard conditions (LB) and following external and internal environment perturbations. The external environment was altered through the addition of glucose to LB media (LBg). All strains were grown in 50 mL of media and OD_{600} readings were taken every half hour for the first 11 hours, then once at 24 hours. All log_2FC values have been normalized to fit between -3 and 3.

After 21 rounds of genome reduction, we assessed the growth of all ten SLIM strains compared to the R_0 control in various media in small volume growth experiments. In LB, seven strains (L1, L2, L4–L7, L9) displayed growth dynamics similar to R_0, with minor differences in lag phase, doubling time, and saturation density (**Figure 2.18a**). In contrast, L3, L8, and L10 exhibited longer doubling times and substantially reduced saturation densities. In TB medium, all strains showed an extended lag phase compared to R_0 (**Figure 2.18b**), but the growth deficits seen in L3, L8, and L10 were largely alleviated; most notably, L7 achieved a higher saturation density than R_0.

To explore the effect of external perturbation, we grew L7 and L8 alongside R_0 and WT in 50 mL LB cultures. These bulk culture experiments replicated previous phenotypes: L7 exhibited a slightly longer lag phase and doubling time than R_0 and WT, while L8 had substantially delayed growth and lower saturation density as compared to L7, R_0 and WT. However, when LB was supplemented with glucose (LBg), both L7 and L8 experienced enhanced growth. Their lag phases and doubling times were shortened, and saturation densities increased to levels exceeding both R_0 and WT (**Figure 2.18d**). This effect was especially pronounced for L8 (**Figure 2.10c**).

Transcriptomic and proteomic analysis of L8 revealed broad downregulation of pathways regulated by the cyclic-AMP(cAMP)/CRP axis, a key regulator of glucose starvation and alternative carbon metabolism[73] (**Figure 2.10b, cyclicAMP-CRP Regulated**). This downregulation was apparent irrespective of the extent of deletions sustained by these pathways. Additionally, L8 showed significant upregulation of rspAB (**Figure 2.19a**), regulators of stationary phase that are activated by cAMP/CRP and promote growth arrest[74]. The combination of

suppressed growth-related pathways and premature stationary phase entry likely explains the growth defect of L8 in LB. Supplementation with glucose alleviates this defect by bypassing CRP-mediated signaling, rescuing L8 growth in LBg.

These findings demonstrate a strong interplay between environmental conditions and genome reduction. Deletion of endogenous DNA dramatically alters the expression patterns of remaining genes. These changes in expression directly affect cellular fitness. However, these expression-mediated fitness changes are highly context-dependent and can be reversed or exacerbated by environmental perturbations. This emphasizes the need to consider environmental context when engineering or utilizing genome-reduced organisms.

SLIM: A modular method for genome reduction of diverse bacteria

The central goal of all genome reduction studies is to distill an organism's genome to its most essential components, a minimal genome comprising only the most fundamental genetic elements[28,75]. In this study, we demonstrated that even within a single species, numerous local minima exist in the space of fully reduced genomes. Moreover, we showed that the process of genome reduction is intimately linked with the cellular environment, creating a dynamic, interdependent system. As a result, generating a library of *E. coli* strains with distinct minimal genomes is likely to uncover a broader set of foundational genetic elements than any single minimal strain. Extending this approach to diverse environmental conditions will likely yield even more unique configurations of minimal genomes. A series of genome minimization experiments conducted across varied contexts will thus help define a more complete and more flexible toolkit of core genetic components for constructing customized *E. coli* chassis.

However, the long-term ambition of genome minimization is not limited to refining a single species. Instead, the broader vision involves constructing truly novel organisms, which requires identifying the fundamental genomic elements of a wide array of phylogenetically distinct microbes. Achieving this goal demands the

development and deployment of genome reduction methods that are generalizable across species (**Figure 2.11b**). To date, no previously published genome reduction platform has demonstrated robust utility beyond a single organism.

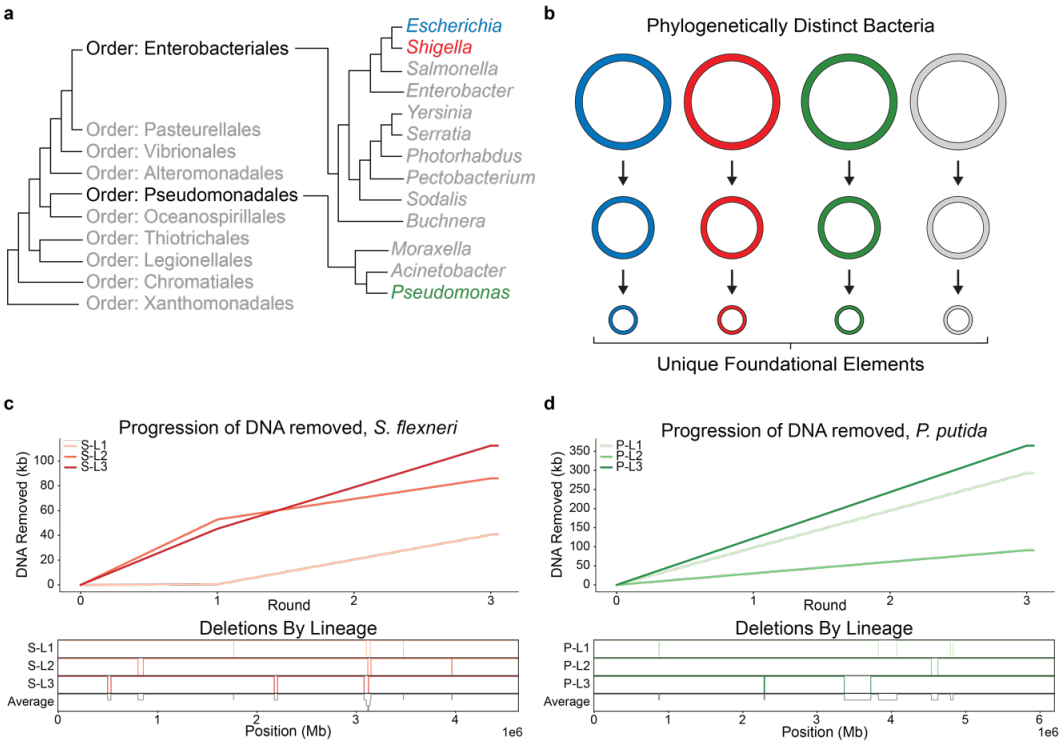

Figure 2.11: Deployment of SLIM for genome reduction across phylogenetic orders
(a) Phylogenetic map depicting the order and genus of bacteria in which SLIM was deployed for genome reduction. (b) Schematic representation of proposed usage of SLIM; to reduce the genomes of numerous, distinct bacteria to their foundational elements. (c) Cumulative DNA removed by round and DNA removed by genomic location following three rounds of SLIM in *Shigella flexneri*. (d) Cumulative DNA removed by round and DNA removed by genomic location following three rounds of SLIM in *Pseudomonas putida*.

To test the modularity and broad applicability of SLIM, we deployed the method in two additional bacterial species spanning different phylogenetic orders: *Shigella flexneri* and *Pseudomonas putida* (**Figure 2.11a**). In *S. flexneri*, we used the standard SLIM workflow to generate a library of three genome-reduced strains across three iterative rounds. After the third round, two clones per strain were subjected to whole genome sequencing. Each strain contained three deletions,

ranging in size from 298 bp to 53 kb (**Appendix A**), with total DNA removed per strain ranging from 37 kb to 112 kb (**Figure 2.11c**).

For *P. putida*, we implemented a slightly modified SLIM workflow due to species-specific differences in counterselection. Specifically, the $pheS^{mut}$/4-chlorophenylalanine system is nonfunctional in *P. putida*, necessitating Cas3-mediated DNA removal without negative selection. While this approach is not viable in *E. coli* (**Figure 2.3b**), the high activity of the Cas3-Cascade system in Pseudomonas species enabled efficient deletion without counterselection. After three rounds of SLIM, an average of two deletions per strain were observed, ranging from 8.2 kb to 356 kb. Total DNA removed per strain ranged from 88.4 kb to 365 kb (**Figure 2.11d**), with the latter representing nearly 6% of the total *P. putida* genome.

Importantly, because SLIM had already been fully developed and optimized in *E. coli*, its application in *S. flexneri* and *P. putida* was rapid, with genome reduction of each species completed in parallel in less than one month. These results highlight SLIM's versatility and efficiency, demonstrating its potential as a broadly applicable platform for unbiased genome reduction across diverse bacteria.

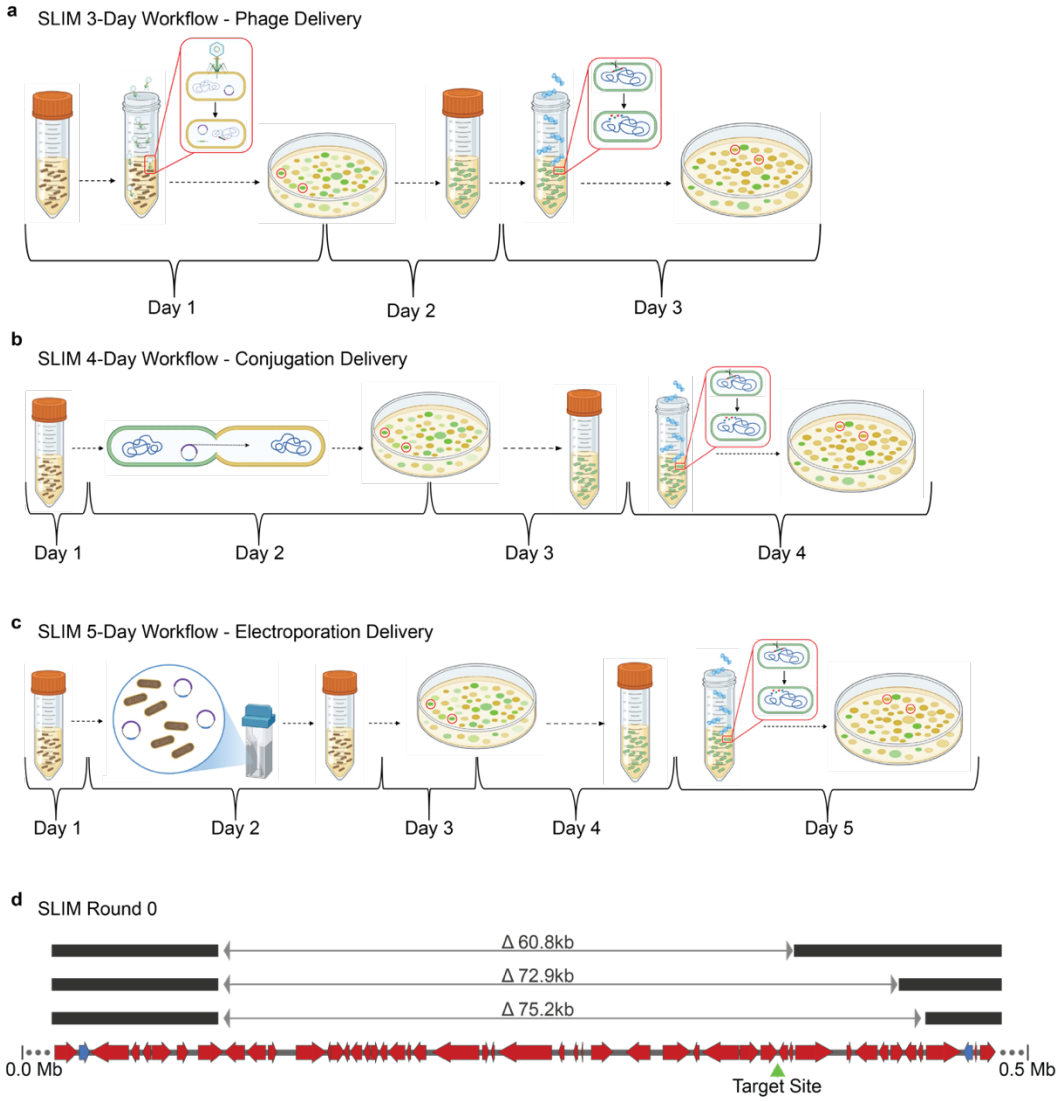

Figure 2.12: SLIM workflows and Round 0

(a) 3-Day workflow established for round 21 of SLIM in E. coli. This workflow utilizes phage transduction for delivery of pTn5. On day 1, colonies are selected from a streak-out plate or a plate from the end of the previous round of SLIM and grown out prior to phage infection. Following pTn5 phage infection, the culture is plated. On day 2, GFP+ colonies, indicating the presence of the target cassette, are selected and grown for cas3 induction. On day 3, cas3 is induced and the culture is plated, selecting for loss of the target cassette. GFP- colonies can then be selected to begin the next round. (b) 4-Day workflow utilizing conjugation for delivery of pTn5. On day 1, colonies are selected from a streak-out plate or a plate from the end of the previous round of SLIM and grown out. On day 2, pTn5 is conjugated from the donor strain to target strain(s). Cells are plated following conjugation and recovery. On day 3, GFP+ colonies, indicating the presence of the target cassette, are selected and grown for cas3 induction. On day 4, cas3 is induced and the culture is plated, selecting for loss of the target cassette. GFP- colonies can then be selected to begin the next round. (c) 5-Day workflow utilizing electroporation for delivery of pTn5. On day 1, colonies are selected from a streak-out plate or a plate from the end of the previous round of SLIM and grown out. On day 2, target strains are made competent and pTn5 is electroporated. Following electroporation, cells are recovered overnight. Following overnight recovery, cells are plated on day 3. On day 4, GFP+ colonies, indicating the presence of the target cassette, are selected and grown for cas3 induction.

On day 5, cas3 is induced and the culture is plated, selecting for loss of the target cassette. GFP- colonies can then be selected to begin the next round. (d) Schematic depicting the genomic location in which the modified target cassette ('Target Site,' green arrow) was integrated prior to Cas3 activation to generate the progenitor strain (R_0) in round 0. The ranges and lengths of deletions observed following Cas3 activation and WGS are indicated. Non-essential genes are colored in red and essential genes are colored in blue. The genomic coordinates of this segment are located between 0 Mb and 0.5 Mb.

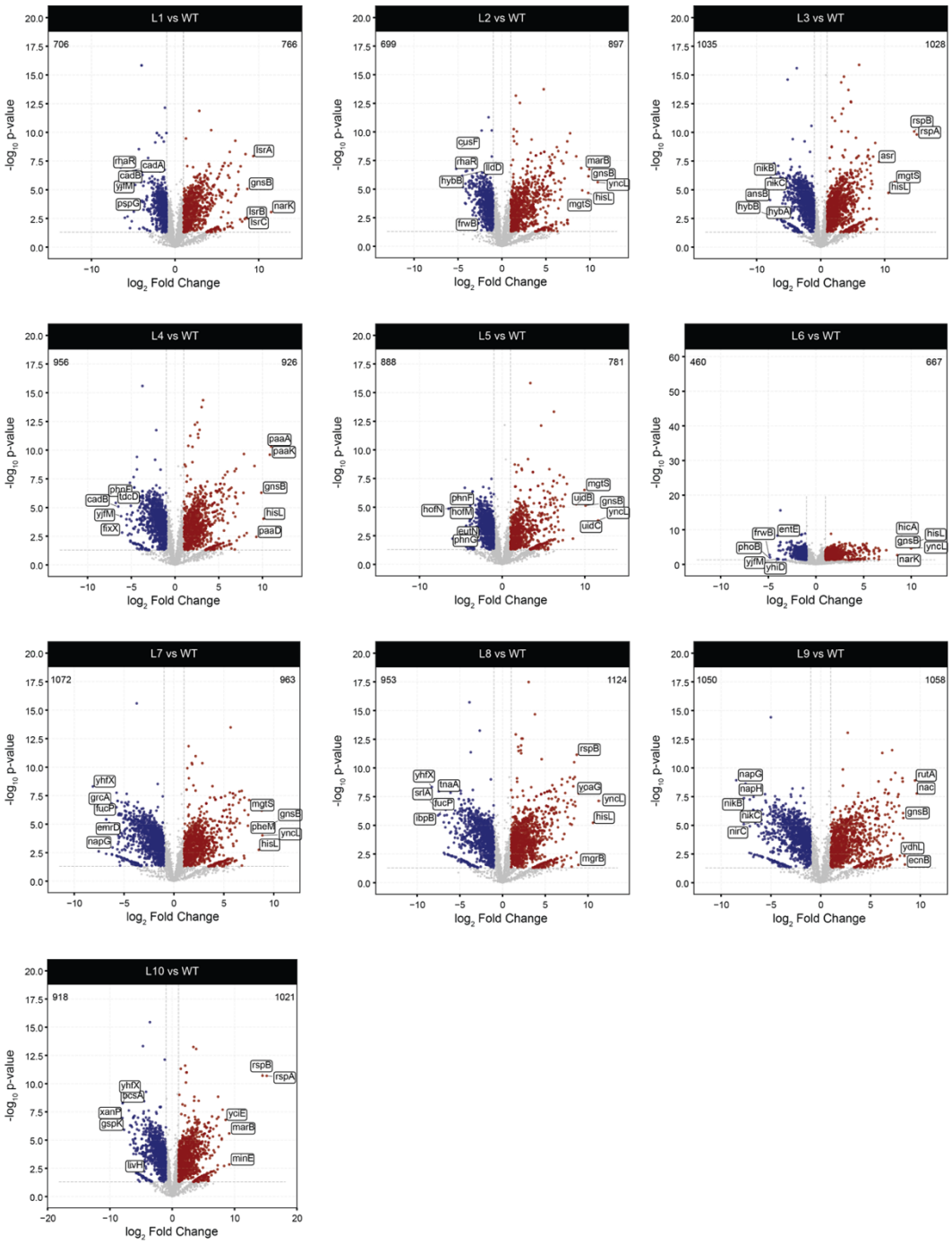

Figure 2.13: RNA differential expression, by genome-reduced strain
Volcano plots depicting differential expression of RNA transcript counts for each genome-reduced strain, as compared to WT. Transcripts were determined be upregulated (red) if $\log_2FC > 1$ and p-value > 0.05 and downregulated (blue) if $\log_2FC < -1$ and p-value > 0.05. All others were designated insignificant (gray). The names of the genes coding for the top 5 most up- and downregulated transcripts, per genome-reduced strain, are flagged and indicated on each plot. The total numbers of

upregulated (top right) and downregulated (top left) transcripts are indicated for each genome-reduced strain.

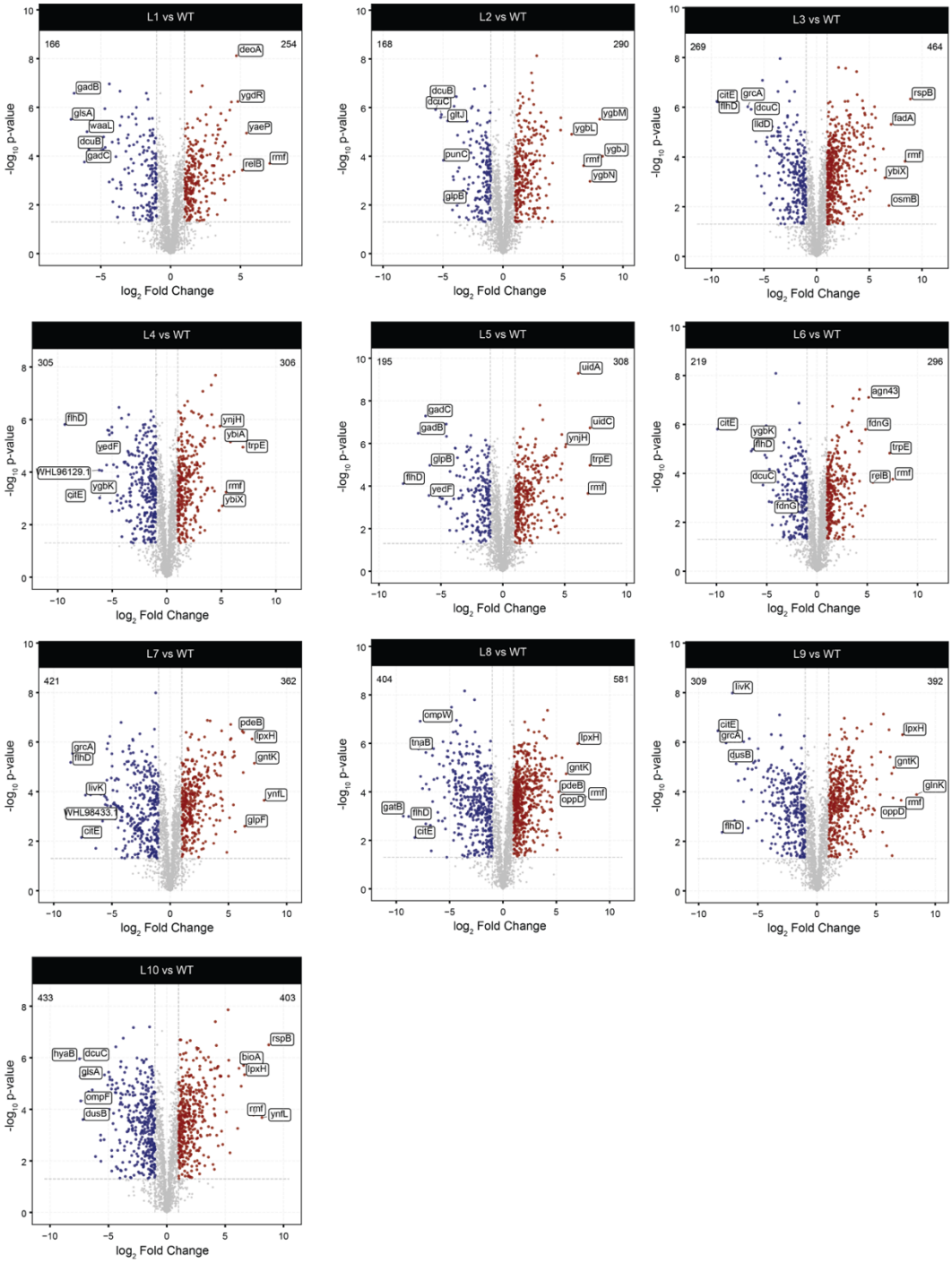

Figure 2.14: Protein differential expression, by genome-reduced strain
Volcano plots depicting differential expression of protein counts for each genome-reduced strain, as compared to WT. Proteins were determined be upregulated (red) if $\log_2FC > 1$ and p-value > 0.05 and downregulated (blue) if $\log_2FC < -1$ and p-value > 0.05. All others were designated insignificant (gray). The names of the genes coding for the top 5 most up- and downregulated proteins, per genome-reduced strain, are flagged and indicated on each plot. The total numbers of upregulated (top right) and downregulated (top left) proteins are indicated for each genome-reduced strain.

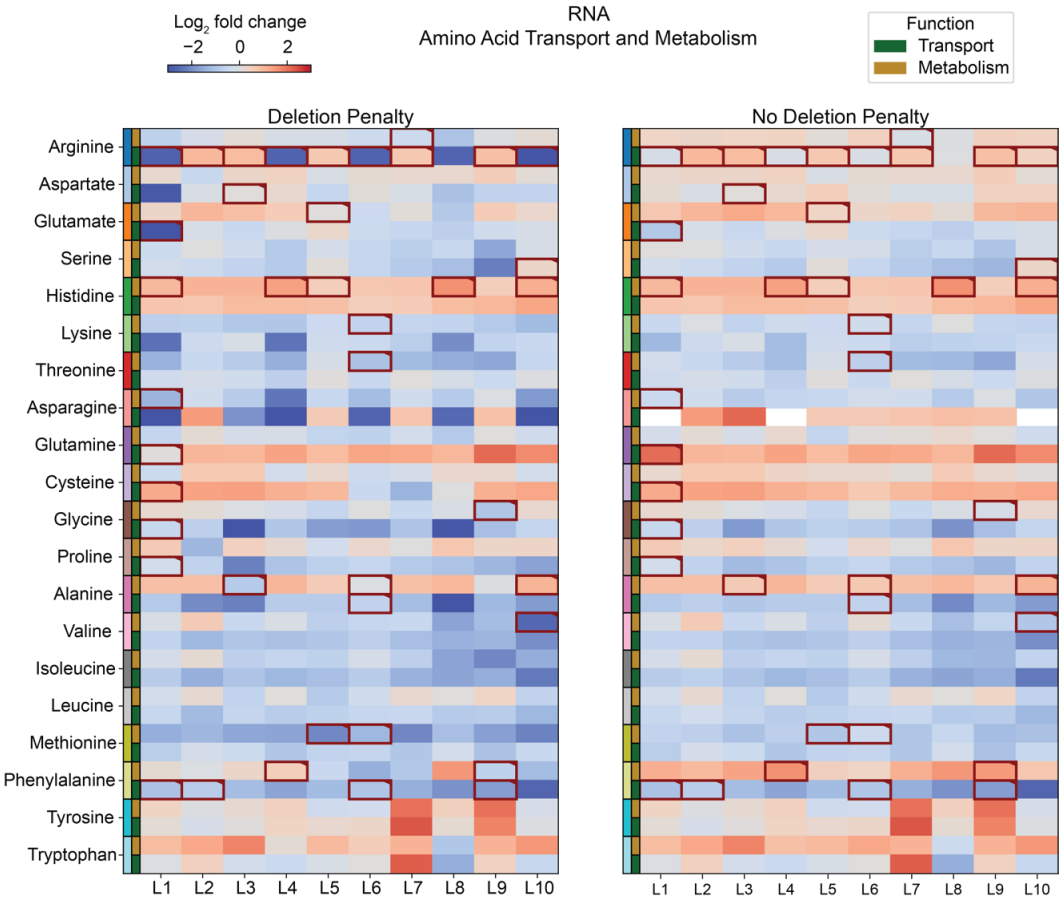

Figure 2.15: RNA amino acid transport and metabolism pathway expression analyses
Heatmaps depicting the extent of differential expression for all amino acid transport and metabolism pathways, as total pathway expression, for each genome-reduced strain as compared to WT. The temperatures indicate the extent of upregulation (red) or downregulation (blue) for each pathway as compared to WT, according to RNA sequencing output. Parallel outputs are displayed for analyses conducted with (left) and without (right) a deletion penalty. All log_2FC values have been normalized to fit between -3 and 3.

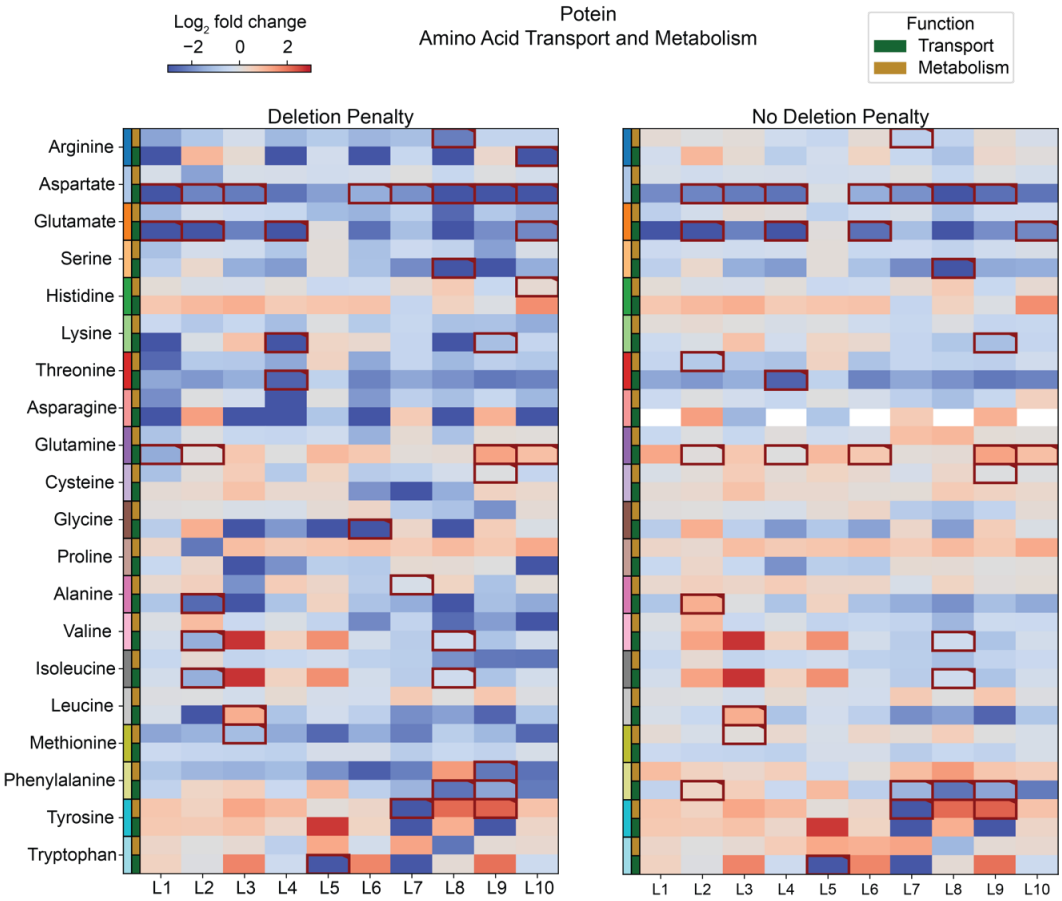

Figure 2.16: Protein amino acid transport and metabolism pathway expression analyses
Heatmaps depicting the extent of differential expression for all amino acid transport and metabolism pathways, as total pathway expression, for each genome-reduced strain as compared to WT. The temperatures indicate the extent of upregulation (red) or downregulation (blue) for each pathway as compared to WT, according to protein sequencing output. Parallel outputs are displayed for analyses conducted with (left) and without (right) a deletion penalty. All log_2FC values have been normalized to fit between -3 and 3.

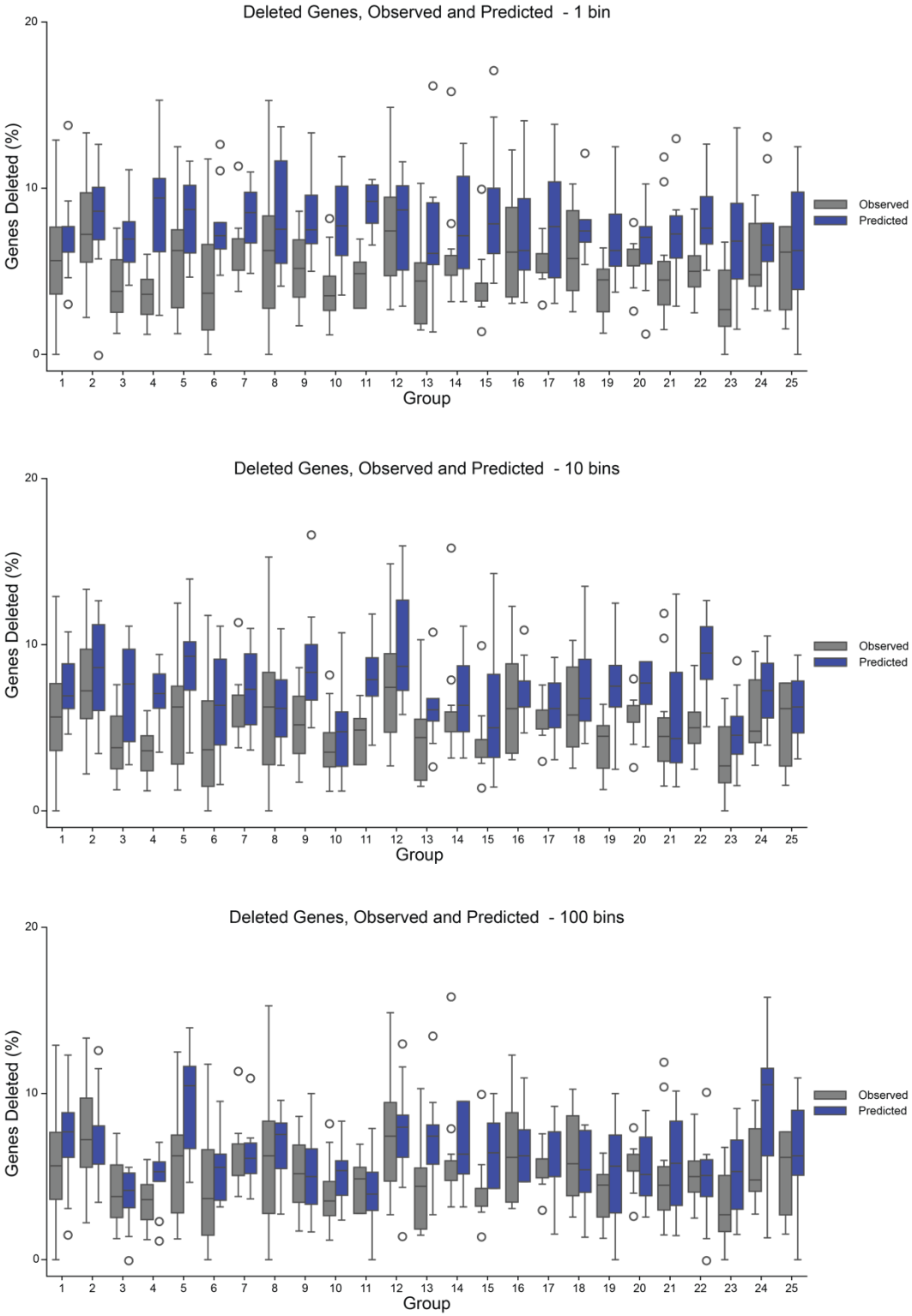

Figure 2.17: Predicted vs observed deletions, varying bin number
Categorical box plots comparing the percentage of observed deletions (gray) amongst all 10 genome-reduced *E. coli* for each of the 25 randomly generated groups to the percentage of predicted deletions (blue) compiled from 10 simulations. Each plot was generated based on data from

simulations conducted using different bin numbers to segment the genome for deletion probability computations; top – 1 bin, middle – 10 bins and bottom – 100 bins.

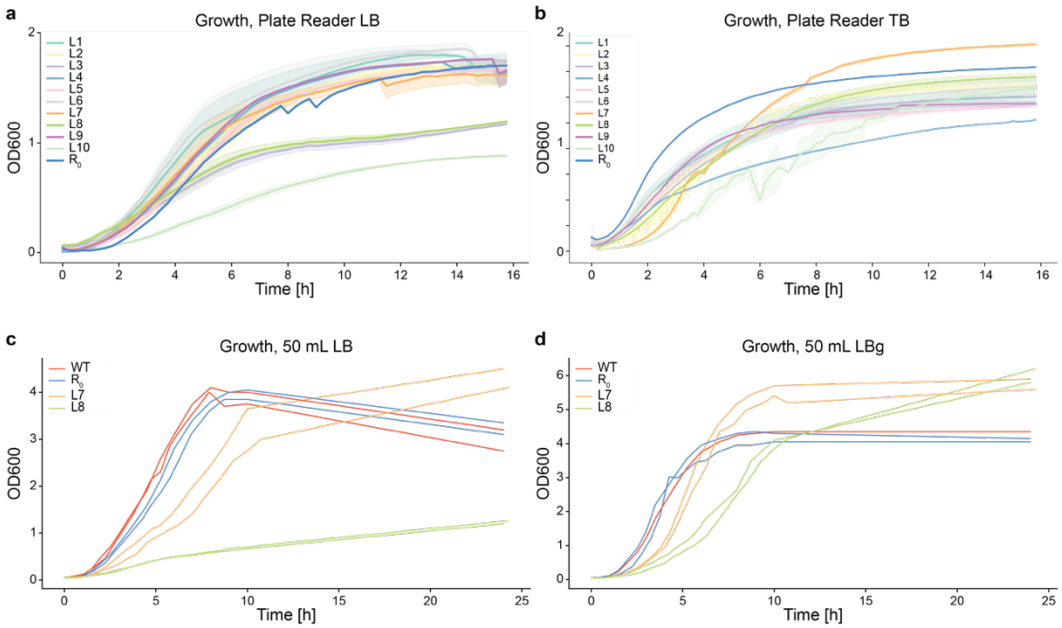

Figure 2.18: Fitness assays with various environmental perturbations
(a) Growth curves for all genome-reduced strains and R_0. All strains were grown in 220 uL of LB media in a 96-well plate. OD_{600} readings were taken every 15 minutes. (b) Growth curves for all genome-reduced strains and R_0. All strains were grown in 220 uL of TB media in a 96-well plate. OD_{600} readings were taken every 15 minutes. (c) Growth curves for WT, R0, L7 and L8. All strains were grown in 50 mL of LB. OD_{600} readings were taken every 30 minutes for the first 11 hours and once at 24 hours. (d) Growth curves for WT, R0, L7 and L8. All strains were grown in 50 mL of LB + glucose (LBg). OD_{600} readings were taken every 30 minutes for the first 11 hours and once at 24 hours.

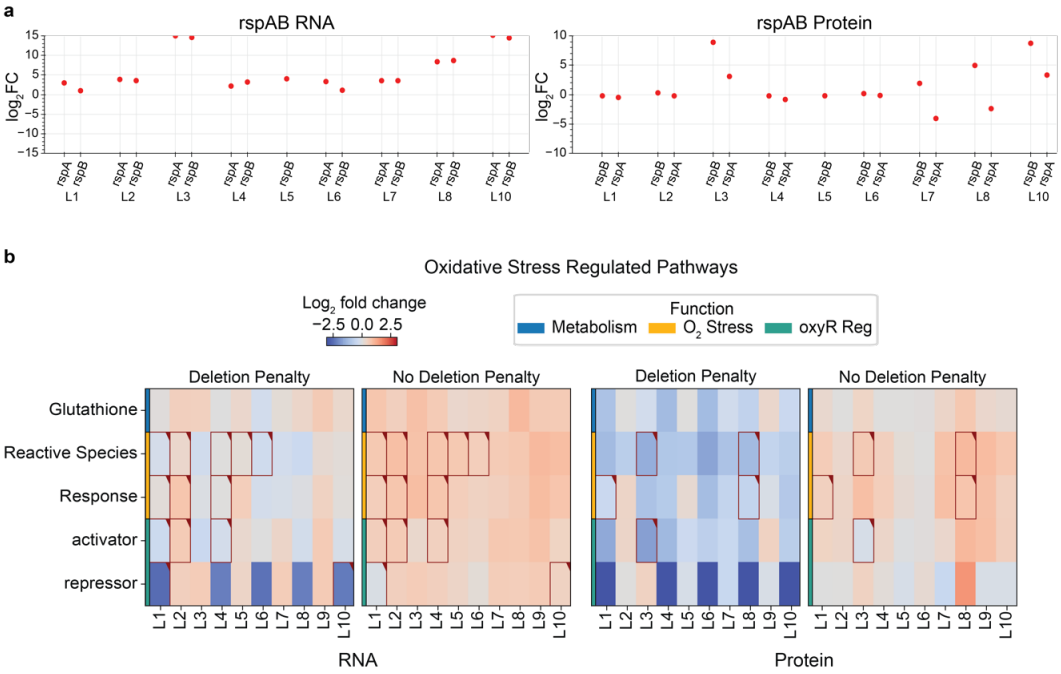

Figure 2.19: rspAB and oxidative stress pathway differential expression analysis
(a) Log_2FC for rspAB transcripts (right) and proteins (left), as compared to WT, for each genome-reduced strain. (b) Heatmaps depicting the extent of differential expression for oxidative stress regulated pathways, as total pathway expression, for each genome-reduced strain as compared to WT. The temperatures indicate the extent of upregulation (red) or downregulation (blue) for each pathway as compared to WT, according to either RNA (left cluster) or protein (right cluster) sequencing output. Within each cluster, parallel outputs are displayed for analyses conducted with (left) and without (right) a deletion penalty. Pathway identifiers are indicated on heatmaps, and functional information is indicated by colored rectangles following pathway identifiers. All log_2FC values have been normalized to fit between -3 and 3.

DISCUSSION

The genomes that we encounter today contain a combination of universally conserved functions, lineage-specific adaptations, and dispensable genetic elements, or genomic "baggage," accumulated and refined over billions of years. The goal of genome minimization is to identify and remove genomic baggage, and other contextually unnecessary genetic components.

However, prior genome reduction efforts have produced only a single or a few closely related genome-reduced strains. These efforts explore only a narrow portion of the possible space of minimized genomes; a conceptual space encompassing all viable genome configurations that arise through different combinations and orders of deletions. Importantly, as deletions are introduced, the expression patterns and regulatory dynamics of the remaining genome change, leading to feedback effects that shape future deletions. Thus, the path taken through genome reduction directly affects the final minimized state, meaning that many distinct minimal genomes are possible even for the same organism, and especially amongst distinct organisms.

Each of these locally minimized genomes, viable endpoints with reduced complexity, offers a unique window into the fundamental structure of genetic networks. Comparative analysis across diverse minimized genomes can reveal previously hidden gene functions, epistatic interactions, and organizational principles. Ultimately, identifying foundational building blocks associated with such functions, interactions and principles will support the construction of universal genomic chassis, laying the groundwork for designing truly novel synthetic organisms. In much the same way that natural evolution has explored diverse genome architectures across billions of years, our aim in this study was to expand synthetic exploration across the full space of minimized genomes.

SLIM is efficient, parallelizable and unbiased

In this study we presented the development and deployment of SLIM, an unbiased and efficient method for genome reduction. Previously reported landmark genome reduction studies have, in some cases, taken over a decade to complete[26,29] and resulted in the generation of a single genome reduced strain. In these studies, a substantial portion of time was sunk in repeated returns to the design-build-test cycle due to systematic failures, a direct result of the intrinsic bias associated with the reliance on prior knowledge to identify essential DNA. Using SLIM, we were able to generate a library of genome-reduced strains, removing up to 14% of the genome, in under 4 years, including development. Time spent actively engaged in the genome reduction process using SLIM was substantially less, cumulatively equating to roughly 3 months to complete all 21 rounds.

A key contributing factor to the speed in which SLIM can be used to reduce genomes is the unbiased nature of performing random deletions. All wild-type genomes contain essential and non-essential genes. As several studies have determined, the category of non-essential genes is composed of a conglomerate of genes of different statuses[68]. While some non-essential genes can simply be deleted, others exhibit a noticeable and sometimes highly detrimental effect upon deletion[26]. Synthetic lethal genes are typically pairs of non-essential genes for which the deletion of one member leads to the other member becoming essential[76]. Quasi essential genes are genes that, when deleted, result in an observable fitness defect[26]. It is possible to elucidate genes that meet these characteristics in wild-type genomes[77]. However, tracking these features as genes are being deleted from a genome is completely intractable because each time a deletion is made, the essentiality landscape has the potential to change. Previously non-essential genes can become essential or quasi essential[77]. Due to this ever-changing essentiality landscape, selectively removing specific segments of DNA can often lead to deletions of newly essential and quasi essential genes. Such events can lead to cell death or severe fitness defects, temporarily or permanently halting the genome

reduction process. In contrast, avoiding the deletion of essential genes, synthetic lethal genes and quasi essential genes is intrinsic to the SLIM DNA removal process. In step 1 of SLIM a genomic locus is selected randomly and uniquely for each cell in a pool of cells. Any cells that sustain a detrimental insertion will be filtered out prior to step 2, preventing the propagation of failure. In step 2 of SLIM, DNA is removed to a random extent. Any cells that experience a detrimental deletion will be lost, leaving only cells that can thrive and propagate following the removal of a random genomic segment.

Other methods for unbiased genome reduction have been previously developed[40,41]. These methods take 7-8 days to complete a single round of DNA removal and application of these methods in *E. coli* failed to generate more than a handful of largely overlapping genome reduced strains. In contrast, a single round of SLIM was shown to take as few as 3 days to complete and the method itself was shown to be highly parallelizable, resulting in the generation of 10 substantially distinct genome-reduced strains.

The generation of a library of genome reduced strains provided the opportunity for a bevy of analyses. In this study, such analyses revealed new insights into the genome reduction process, including information on deletion permissibility and global expression dynamics in response to both the genome reduction process in general. Analysis of expression dynamics in the context of deletions allowed us to identify potential gaps in pathway annotations. Finally, interrogation of environment-dependent phenotypic behaviors in the context of expression dynamics allowed us to identify such expression dynamics as the causal link between deletions of some genes and observed changes in fitness.

Deletion permissibility is dictated by multiple factors

Following 21 rounds of genome reduction with SLIM, we explored the factors dictating the removal of specific segments of DNA across all 10 members of the genome-reduced *E. coli* library. In this set of analyses we examined two

characteristics of genes, gene product function and genomic location, in isolation to determine what role, if any, that each of these characteristics plays in determining the deletion permissibility of any single gene. By binning genes into functional groups and assessing deletions amongst these groups across all genome-reduced strains we were able to identify patterns suggesting that genes in some functional groups appear to be more susceptible to deletion than others. Based on quantitative analysis assessing the median percentage of genes deleted, COG groups such as 'Inorganic ion transport and metabolism,' 'Transcription,' and 'Secondary metabolites biosynthesis, transport and catabolism' exhibit a relatively high deletion permissibility while groups such as 'Translation, ribosomal structure, and biogenesis' and 'Replication, recombination, and repair' exhibit comparably lower deletion permissibility. The underlying factors compelling this apparent discrepancy in deletion permissibility amongst functional groups is unclear. It is possible that groups with high deletion permissibility contain more non-essential genes and processes or more redundancies than groups with low deletion permissibility. In addition, we explored the extent to which genomic location plays a role in deletion permissibility. For this analysis we binned the genome by genomic location, computed the average coverage within each bin on a per-nucleotide level and used this metric to assign a deletion probability to every gene contained in each bin. We then simulated deletions based on these deletion probabilities. Following deletion simulations we randomly assigned each gene into one of 25 groups and compared deletion percentages from each group, per simulation, to observed deletion percentages. In this analysis, we observed that most groups show a strong correlation between predicted and observed deletions, with an average R^2 of 0.47. It is important to note that this result was dependent on the number of bins chosen for deletion simulations; fewer bins resulted in lower correlation between observed and predicted deletions. Simulations in which deletion probabilities were randomly assigned to bins resulted in an average R^2 of –2.2, exemplifying the strength of the genomic location-based deletion probabilities model.

The observations made in the analyses of deletions point to a non-random pattern of deletions. Within the deletions data we have found evidence of both essentiality genes and TA pairs bounding deletions. The genomic location-based deletion probabilities model captures roughly half of the variance apparent in observed deletions and analyses of deletions in the context of functional groups does appear to indicate trends of deletion permissibility variation across the 25 COG groups. Taken together, these observations indicate a role for both function and genomic location in the determination of deletion permissibility of any individual gene. However, neither gene function nor genomic location are sufficient to fully explain the deletion patterns that we have observed through 21 rounds of genome reduction. The absence of a truly predictive characteristic or set of characteristics highlights the difficulty in accurately predicting permissible deletions, further lending credence to genome reduction methods that remain agnostic to prior information and such predictions.

For this set of analyses, we attempted to isolate gene product function from genomic location. In reality, function and genomic location are often intrinsically linked. In *E. coli*, all copies of the rrn operon, which encodes 5s, 16s and 23s rRNAs, are located proximal to the origin of replication (oriC) in the top half of the genome. During rapid population expansion, this portion of the genome is replicated more frequently to facilitate more efficient proliferation, resulting in the accumulation of more copies of the rrn operon[78]. The effects of genomic location are not limited to a genes proximity to replication nodes such as the oriC or the replication termination site (ter) as the 3D structure of the genome can result in the colocalization of distant regions[79]; a phenomenon that also has the potential to effect expression. It is likely that deliberate placement of genomic components will be required when building novel, de novo genomes, making a detailed understanding of the geography of natural genomes essential. Although more detailed information is required, the factors dictating deletion permissibility described in this study could help determine the design of such genomes.

Multi-omics analyses following 21 rounds of SLIM in *E. coli* yield novel biological insights

Previous studies have shown that deleting even a single gene can alter the expression of many others[45,46]. Due to the complex network of interactions amongst gene products, the deletion of a single gene will not only result in the complete loss of RNA and protein expression of that gene's product but can also result in changes in expression amongst all related genes. We employed a multi-omics analysis approach to evaluate the library of 10 genome-reduced *E. coli* following 21 rounds of genome reduction with SLIM. We found that the number of transcripts and proteins showing differential expression far exceeded the number of deletions made. Following RNA and protein sequencing, we interrogated RNA transcript count distribution and protein count distribution in the context of 25 functional COG groups. This analysis revealed differences in RNA and protein allocation between genome-reduced strains and WT. RNA transcript analysis revealed that genome-reduced strains allocate a greater fraction of their transcriptional output to genes associated with the 'Chromatin structure and dynamics' group. The three genes that compose this group, hns[61], dps[62] and hupA[63] all perform functions related to DNA protection during stationary phase and in response to various stressors. Both RNA transcript and protein count analyses show that genome-reduced strains allocate fewer resources than WT for both transcription and translation of genes associated with the 'Translation, ribosomal structure and biogenesis' group; a group in which WT allocates a substantial portion of its transcription and translation resources. The emphasis that genome-reduced strains appear to be putting on functions associated with DNA protection and de-emphasis on normal cellular functions suggests that genome-reduced strains exhibit a persistent stress or defensive state. Differential expression analysis between genome-reduced strains and WT, across pathways and amongst individual genes, uncovered additional global responses. Specifically, DE analysis showed that according to both RNA transcript counts and protein counts, genome-reduced strains highly upregulate rmf as compared to WT (**Figure 2.9b**). In fact, rmf ranks among the top five most

upregulated genes at the protein level in several genome-reduced strains (**Figure 2.14**). The rmf protein product is responsible dimerizing 70s ribosomal subunits into inactive 100s ribosomal subunits[68] as the cell enters stationary phase so that, upon transition from stationary phase back to exponential phase, ribosomal subunits are readily available. The overexpression of rmf seems to indicate that genome reduced strains are in a heightened state of preparation for hibernation due to stress induced by DNA removal. Importantly, these effects persist across many generations after deletion, implying they are stable, genome-encoded responses rather than transient. The coordinated and stable upregulation of rmf strongly suggests that genome-reduced strains adopt a hibernation-like state as a long-term adaptation to genomic stress.

When we analyzed differentially expressed pathways across all genome-reduced strains, we found both broad, shared expression shifts and some unexpected strain-specific changes. One especially surprising result emerged from looking at how cells responded to our selection scheme. As mentioned earlier, we use a mutant version of *pheS* (*pheS*^{mut}) in our target cassette to enable counterselection and encourage Cas3 activity. The system is designed so that when *pheS*^{mut} is expressed in the presence of 4-chlorophenylalanine, the analog gets taken up through the phenylalanine transport pathway and is mis-incorporated into proteins, causing misfolding and ultimately cell death[55]. Under this pressure, the cell has two obvious survival strategies: (1) allow Cas3 to chew through the *pheS*^{mut}-containing cassette (plus flanking endogenous DNA), resolve the break, and move on or (2) acquire a mutation in *pheS*^{mut} that makes it non-functional, rendering the toxin ineffective. We specifically designed the SLIM workflow to detect and avoid the second option, while promoting the first. After every round, we manually screen survivors to ensure that Cas3 activity actually occurred, and we reintroduce a fresh copy of *pheS*^{mut} at the start of each round to eliminate the chance of propagating resistance through mutational escape. But even with all those safeguards in place, the genome-reduced strains still managed to find an unexpected third way out. According to our

RNA-seq data, all genome-reduced strains consistently downregulated genes involved in the phenylalanine transport pathway. Since this pathway is responsible for importing 4-chlorophenylalanine, turning it down reduces the amount of toxin entering the cell, conferring a selective advantage for genome-reduced strains, without touching *pheS*^{mut} at all. This was completely unexpected and points to a novel mechanism of resistance to *pheS*^{mut}/4-chlorophenylalanine mediated toxicity. This observation reflects the broader stress-adaptive behavior we've seen in many genome-reduced strains; they seem especially adept at mounting long-term responses to selection pressure by rerouting expression in clever ways. More broadly, this result underscores how even tightly controlled selection systems can be circumvented through global shifts in gene regulation, particularly in cells navigating extensive genomic perturbation.

The analysis of a library of genome-reduced strains in the context of multi-omics expression data allows for the association of specific expression patterns with specific deletions. According to both RNA transcript and protein DE, the glutamate transport pathway is downregulated in all genome-reduced strains except for L5. As previously mentioned, glutamate transport is mediated by the gltIJKLS family of genes, which is located in the tandem repeat section of DH10B. The full tandem repeat section consists of two identical and sequential 100+ kb segments of DNA, meaning each gene found in the tandem repeat section will have two copies. All strains except L5 have deleted one copy or most of one copy of the tandem repeat, resulting in the loss of one copy of each gene in the gltIJKLS gene family. This deletion likely explains the widespread reduction in glutamate transport activity amongst all genome-reduced strains except for L5. A parallel pattern of expression can be observed for threonine metabolism. No known threonine metabolism genes are located in this region, suggesting the involvement of unannotated regulators or long-range genetic interactions. annotated. This result highlights the value of genome reduction as a tool to uncover gaps in current functional annotation. Even with a relatively small library of 10 genome-reduced strains, we identified specific,

mechanistically plausible links between deletions and expression phenotypes. Scaling this approach could reveal many more such relationships.

Wholesale changes in expression, combined with environmental perturbations, result in substantial fitness changes

Previous work has defined "quasi-essential" genes as those whose deletion significantly impairs cellular fitness without causing lethality[26]. Building on this concept, we sought to elucidate the causal link between the deletion of quasi-essential genes and changes in cellular fitness. In this study, we have observed that deletions lead to wholesale changes in both RNA and protein expression of remaining genes in all genome-reduced strains. In addition to expression analyses, we examined the growth behaviors of all genome-reduced strains and R_0 in LB. Growth curve analysis in LB revealed marked fitness reductions in strains L3, L8, and L10 compared to WT and R_0. To explore the source of these fitness reductions in more detail we conducted further growth behavior analyses with a subset of genomes reduced strains, L7 and L8, in addition to WT and R_0, in a variety of environments. To probe the context-dependence of these defects, we altered the external environment by supplementing LB with glucose (LBg). Upon alteration of this environment, we observed substantial changes in growth behaviors for one of the genome-reduced strains, L8. Based on these observations, it is clear that changes in expression, induced by deletions, combined with changes in environment, can substantially alter cellular fitness. Transcriptomic and proteomic data from L8 revealed widespread downregulation of genes regulated by the cAMP/CRP axis, a key transcriptional program activated during glucose starvation. When glucose is added to LB (LBg) the reliance on functions regulated by the cAMP/CRP axis is relieved and a substantially improved fitness for L8 is observed. In comparison to growth behaviors in LB, when L8 is grown in LBg it exhibits a clear and robust exponential phase and grows to a substantially higher maximum density, even exceeding that of WT. These observations provide a direct link between deletions and changes in fitness, strongly suggesting that expression

changes induced by deletion events, rather than just the deletions themselves, are the primary drivers of altered fitness states.

Throughout this study we have been adamant in avoiding the attribution of any expression or phenotypic changes amongst genome-reduced strains to the deletion of a single gene, as any observed differences are undoubtedly the cumulative result of numerous deleted genes, whether within a single round or across all 21 rounds of SLIM-mediated genome reduction. However, there is one instance in which a compelling case can be made, as the CRP gene has been deleted from L8. In a recent study evaluating a CRP single deletion mutant *E. coli*, it was shown that the deletion of CRP results in an almost 50% decrease in growth rate in glucose-free media and the differential expression of over 700 transcripts. It is likely that the deletion of CRP from L8 served as a major catalyst for the expression differences and dramatic changes in growth phenotype observed following environmental perturbations[47].

This work highlights the importance of environmental context when conducting and evaluating the results of the genome-reduction process. The shifting growth behaviors of L8 reveal that different environments emphasize different genetic dependencies, suggesting that the concept of a "minimal genome" is not absolute, but conditional. This phenomenon was further supported by RNA and protein expression data. Future genome-reduction efforts in diverse conditions could reveal a broader landscape of viable minimized genomes, each tailored to its environmental niche, revealing distinct foundational elements. The elucidation of foundation genomic elements for bacteria cultured in distinct environments will have major implications on design considerations for novel genomic architectures.

SLIM exhibits modularity and efficacy in the genome reduction of multiple, phylogenetically distant, bacteria

All genome-reduction methods to date have been limited to a single bacterial species, constraining both the diversity of minimized genomes produced and the ability to identify universally essential functions. SLIM's species-agnostic

architecture enabled us to address this limitation directly by applying it across phylogenetically and physiologically distinct bacterial systems. We performed three rounds of SLIM in each of *S. flexneri* and *P. putida*, generating libraries of three genome-reduced strains per species. Remarkably, the transition from *E. coli* to these organisms required no additional optimization. The parallelizability and modularity of SLIM allowed us to generate these libraries in just three weeks, a stark contrast to the years often required by traditional bottom-up or directed top-down minimization techniques. Notably, within this short timeframe, we removed nearly 6% of the wild-type genome in the genome-reduced *P. putida* strain P-L3. These results underscore SLIM's modularity and broad applicability, suggesting that its components can be readily adapted for rapid, high-throughput genome reduction across a wide range of bacterial species.

CONCLUSION AND FUTURE DIRECTIONS

Conclusion

In this study, we present SLIM as a highly efficient, modular, and parallelizable platform for unbiased genome reduction across phylogenetically diverse bacteria. Unlike previous genome-reduction methods that focus on generating a single or few largely overlapping minimized strains, SLIM enables the creation of libraries of genome-reduced strains, greatly expanding analytical power. By comparing multiple genome-reduced strains to wild type, we were able to capture both shared and strain-specific transcriptional and translational responses to large-scale DNA removal. These comparisons revealed a global systems-level response to genome reduction, uncovered an unexpected mechanism of counterselection escape, and highlighted specific pathways altered across minimized strains, suggesting new opportunities for biological discovery.

Environmental perturbation experiments further demonstrated that the fitness of genome-reduced strains is highly dependent on environmental context. These findings underscore that the functional consequences of genome reduction are not solely determined by which genes are deleted, but by how the remaining genome responds—transcriptionally, translationally, and physiologically—to both the deletions and the environment. In one such perturbation, we demonstrated that a genome-reduced strain can outperform wild type in specific conditions when paired with a minimal genetic addition, illustrating the potential of minimized genomes as adaptive and customizable chassis for future synthetic biology applications.

Altogether, this work highlights the analytical power and practical promise of generating diverse libraries of genome-reduced strains. The space of all possible minimized genomes is vast: for any given species, there exists a large number of viable genome configurations, each reflecting a unique sequence of deletions and adaptive responses. Expanding into this space is essential for uncovering fundamental building blocks of cellular life. These foundational components, once

identified, can be used to construct versatile, rationally designed genomic chassis. Until now, the generation of genome-reduced strains has been a slow and manually intensive process, often taking years and yielding a limited view of this broader landscape. SLIM overcomes these limitations. It is the first genome reduction platform capable of producing minimized strain libraries across multiple species, within weeks, and without species-specific re-optimization. The throughput and generalizability of SLIM bring us substantially closer to the long-standing goal of de novo synthesis of novel, fully engineered organisms.

Future Directions

The work presented in this study represents only the beginning of what SLIM can enable. Moving forward, we are focused on evolving the SLIM workflow in both scale and efficiency. Initial proof-of-concept experiments have demonstrated the feasibility of reducing the liquid culture portions of the SLIM workflow to small-volume formats. In round 20 of genome reduction in *E. coli*, we successfully carried out the liquid culture portions of steps 1 and 2 in a 96-well plate format for the remaining 9 library members (reduction of L5 was halted at round 7). This transition, completed in just 4 days, indicates the potential to scale up the size of the libraries generated with SLIM nearly tenfold.

Beyond scaling, we are working toward automating and streamlining the workflow. The current protocol includes two plating steps; one after step 1 and another after step 2. We have developed initial designs for a fully liquid-phase version of SLIM that would eliminate these intermediate plating steps, allowing for continuous rounds of genome reduction. Such a system could be executed entirely by a liquid handler, minimizing human error and dramatically increasing throughput.

Further, we are developing a compressed, single-step version of the SLIM workflow. This would reduce the time per round to just two days. Initial tests have shown early promise, suggesting that future iterations of SLIM may be faster and more scalable than ever before.

Alongside these technical improvements, we anticipate broader adoption of SLIM by the synthetic biology and microbiology communities. The protocols and results presented here serve as a blueprint that other groups can adapt to reduce genomes in a wide variety of bacterial species. As we demonstrated in *E. coli*, genome reduction in other organisms is likely to reveal novel biology and new foundational building blocks. The accumulation of such data across diverse bacteria and conditions will generate a rich toolkit for constructing novel, modular genomic chassis.

A central goal of genome reduction is to engineer strains with streamlined, task-specific genomes optimized for performance in defined environments. While this vision is still emerging, several genome-reduced strains have already shown promise in industrial applications. For example, genome-reduced *E. coli* strains MDS42 and MGF-01 have outperformed their wild-type counterparts in L-threonine production[32,80]. Similarly, a genome-reduced *E. coli* strain generated using TMRD has shown enhanced polyhydroxybutyrate production[81], and the genome-reduced *Bacillus subtilis* strain PG10 exhibited improved recombinant protein production compared to its wild-type ancestor, *B. subtilis* 168[82].

The natural next step for the field is to extend genome reduction into eukaryotic systems. Several early-stage methods suggest this is feasible. The SCRaMbLE system (Synthetic chromosome rearrangement and modification by loxP-mediated evolution) was developed to induce large-scale genomic alterations in yeast. SCRaMbLE not only produced rearrangements, inversions, and duplications, but also generated deletions in synthetic yeast chromosomes[83]. More recently, prime editing strategies like PEDAR and PRIME-del have achieved modest genomic deletions in mammalian cells[84,85].

Although eukaryotic genome reduction remains in its infancy, a eukaryotic adaptation of SLIM could accelerate progress significantly. Key components of SLIM already have analogs in eukaryotic systems. Transposon systems like

piggyBac[86] and Sleeping Beauty[87] have been shown to function efficiently in eukaryotic cells, although they currently display stronger sequence insertion preferences than Tn5. Emerging variants, such as the hyperactive Sleeping Beauty[88] and the Mage transposase[89], offer increased activity and reduced bias, making them promising candidates for random insertion in future adaptations of SLIM.

Additionally, a Type I-E CRISPR/Cas3 system, closely related to the Type I-C system used in SLIM, has been shown to induce deletions in mammalian genomes[90]. Taken together, these elements provide a strong foundation for the development of a SLIM-like genome reduction system in eukaryotic organisms.

In sum, SLIM opens the door to a new era of high-throughput, parallel genome reduction across bacterial species and potentially beyond. As we continue to iterate on its design and expand its scope, we move closer to the long-term goal of rationally engineering streamlined, robust organisms from the ground up.

MATERIALS, METHODS AND DATA AVAILABILITY

General Methods

DNA amplification was performed using PrimeSTAR™ GXL DNA Polymerase (TaKaRa Bio USA cat # R050) unless otherwise stated. DNA oligonucleotides were obtained from Integrated DNA Technologies (IDT) and Millipore Sigma. Nucleic acid purification and cleanup was done with QIAquick® PCR Purification Kit (QIAGEN™ cat # 28106). DNA extraction from agarose gels was conducted with GeneJET Gel Extraction Kit (Thermo Scientific™ cat # K0691). All plasmids were assembled using Gibson assembly using NEBuilder® 2x HiFi Assembly mix (NEB cat # E2621). Standard plasmid DNA cloning was performed using electrocompetent E. coli DH10B, which was made electrocompetent as previously described[91]. Plasmid DNA extraction from E. coli DH10B was done with QIAprep® Spin Miniprep Kit (QIAGEN™ cat # 27106). All antibiotic and counter-selection concentrations when used are as follows: 100 μg/ml streptomycin, 250 μg/ml hygromycin, 5 mM 4-chlorophenylalanine, 25 μg/ml chloramphenicol, 7.5% w/v sucrose, 50 μg/ml kanamycin, 100 μg/ml carbenicillin, and 15 μg/ml gentamicin

Bacteria Culturing Conditions

All *E. coli* and *S. flexneri* bacterial strains were grown at 37°C in selective Luria-Bertani (LB) liquid media with shaking unless otherwise stated. Individual colonies were grown overnight on selective LB-agar plates at 37°C prior to liquid growth. All *P. putida* bacterial strains were grown at 30°C in LB liquid media with shaking or on solid LB-agar plates. *E. coli* DH10B rpsL K43R, *Pseudomonas putida* KT2440 and *Shigella flexneri* CFS100 rpsL K43R ΔrecA Δtus 1201 ΔoriT ΔhigB Δhok were used as parental strains for genome reduction. *E. coli* DH10B rpsL K43R was used as the host strain for all plasmid DNA cloning and as the base for

genRK24 construction. *E. coli* C600 P1kcΔcoi::kanR was used as the base for p1kc_piRsC construction.

Plasmid Construction

All plasmid constructs described in this section were constructed via Gibson assembly, unless otherwise specified. In short, pTn5-v2 was constructed using components from various sources. The target cassette containing the Cas3/pdeL target sequence, $pheS^{mut}$, HygR, and GFP was constructed with pieces taken from unpublished plasmids generated in-house. The pdeL target sequence and Tn5ME were added directly to the cassette during amplification. The pSC101 origin and sacB genes were taken from the same plasmids. The hyperactive Tn5 transposase gene was derived from psfTn5 (Addgene #79107). The araCE operon was derived from the Marionette cassette[92]. To construct pTn5-v5, the pSC101 origin of replication and the sacB gene in pTn5-v2 were replaced with the R6Kγ origin of replication from pDonor_ShCAST_kanR (Addgene #127924). To construct pTn5-v11 phagemid components were added to pTn5-v5 from pBBa_J72113-BBa_J72152 (Addgene #40785). pCas3 was constructed from pCas3cRH (Addgene #133773), with the addition of pdeL specific crRNA. To construct pCas3, a pdeL specific crRNA amplicon was inserted into purified pCas3cRH following digestion with BsaI restriction enzyme (New England Biolabs). An inducible T4 ligase from pCA24N-ligase (Addgene #87741) was added to pCas3 in a sequential assembly step.

Starting strains construction

E. coli DH10B, *S. flexneri* and *P. putida* were all transformed with pCas3 prior to the application of SLIM for genome reduction. *E. coli* was further engineered to produce the progenitor strain before SLIM was used to reduce the *E. coli* genome.

Progenitor strain construction

The progenitor strain, or SLIM starting strain, was created in an effort to remove the wild-type occurrence of the Cas3 target sequence used in both the original and modified target cassettes. This strain was generated from a version of our wild-type strain, *E. coli* DH10B, containing pCas3. First, the modified target cassette was integrated into a location directly downstream of the wild-type location of the target sequence using the Lambda-Red recombination system encoded by pKW20[93]. After curing pKW20, the resulting strain was grown from a streak-out isolate overnight in 10 mL LB + gentamicin and hygromycin. The next day, the culture was pelleted by centrifugation, the supernatant was discarded and the pellet resuspended in fresh LB media containing gentamicin and 0.3% rhamnose. After incubating for 5.5 hours in a shaking incubator at 37°C we plated the cells on LB/agar plates supplemented with gentamicin and 4-chlorophenylalanine, selecting for loss of the target cassette. Three GFP- colonies were selected and screened, eventually selecting colony 1 as the progenitor strain (**Figure 2.11d**).

Genome-reduced strain construction

Genome-reduced lineages 1–10 were generated during the first round of SLIM. Briefly, 20 colonies were selected following step 1 of SLIM during round 1 of SLIM. The colonies were grouped into groups of two to form 10 independent lineages. Genome-reduced lineages 1–10 were further developed throughout each of the 21 rounds of SLIM.

E. coli DH10B + target cassettes construction

Ten strains were generated to test the DNA removal efficacy of the Cas3-cascade system by integrating either the original target cassette or the modified target cassette into one of five genomic locations, independently in a single cassette per strain fashion. The five genomic locations, listed by non-essential span, are as

follows: 8 kb, 56 kb, 78 kb, 90kb and 290 kb. The protocol used to generate each of the strains follows directly from that used to generate the progenitor strain. *E. coli* DH10B was transformed with pCas3 prior to the integration of target cassettes.

E. coli p1kc_piR-sC construction

E. coli p1kc_piR-sC was used to package pTn5-v11 in P1 phage particles. This strain was developed from *E. coli* P1kcΔcoi::kanR, a variant of *E. coli* C600 that contains a P1 bacteriophage genome[94]. Briefly, the wild-type pi replicator (piR) sequence was amplified and assembled with the sacB (s) and chloramphenicol resistance (C) genes. The resulting piR-sC cassette was again amplified and integrated into the P1kcΔcoi::kanR genome in place of kanR using the Lambda-Red recombination system, as described previously.

E. coli piR_genRK24 construction

E. coli piR_genRK24 was used as a donor strain, to deliver pTn5-v5 via conjugation. *E. coli* piR_genRK24 was developed from *E. coli* DH10B_genRK24. *E. coli* DH10B_genRK24 was developed as a high frequency recombination strain in the lab previously (unpublished). To engineer *E. coli* piR_genRK24 from *E. coli* DH10B_genRK24, the same piR-sC cassette used to build *E. coli* p1kc_piR-sC was amplified and integrated into the *E. coli* DH10B_genRK24 genome.

SLIM workflows

The procedures below are described singularly. For the generation of the ten-member SLIM library, all steps described below were completed for each member of the library in parallel.

pTn5 delivery/transposition, phage transduction, 3-day

Day 1: To begin each round, two colonies of SLIM target cells were picked from plates, either streak outs of glycerol stocks from previously picked colonies or colonies on plates from the end of the previous round of SLIM, and resuspended in 30 uL 10% LB. The resuspensions were used to independently inoculate 1.5 mL ePLM (LB 140 mM MgCl2 and 7 mM $CaCl_2$) + gentamicin (15 ug/mL) + 0.5% arabinose (to pre-induce cells for Tn5 transposition) cultures in a 13 mL culture tube which were then incubated in a 37°C shaker for 4 hours at 170 rpm. Next, the independent cultures were combined and added to 1 mL of pTn5v11 phage lysate; the resulting culture is incubated in a 37°C shaker for 20 minutes. The culture was then incubated statically at 37°C for 20 minutes. The phage infection reaction is then quenched with 1mL of SOC supplemented with 200 mM sodium citrate and the culture was incubated in a 37°C shaker for 40 minutes. Cells were then pelleted by centrifugation and resuspended in 10 mL LB + gentamicin + hygromycin (50 ug/mL) + 0.5% arabinose and incubated in a 37°C shaker for 3 hours. Cells were then pelleted by centrifugation and resuspended for plating; all cells were plated on LB/agar plates supplemented with gentamicin and hygromycin, in full and as a dilution series.

pTn5 delivery/transposition, phage transduction, 4-day

Day 1: To begin each round, two colonies of SLIM target cells were picked from plates, either streak-outs of glycerol stocks from previously picked colonies or colonies on plates from the end of the previous round of SLIM, and resuspended in 30 uL 10% LB. The resuspensions were used to independently inoculate 3 mL LB + gentamicin + 0.5% arabinose (to pre-induce cells for Tn5 transposition) cultures which are incubated in a 37°C shaker overnight.

Day 2: The next day, 750 uL from each culture was combined, cells were pelleted by centrifugation and resuspended in 1.5 mL ePLM + gentamicin (15 ug/mL) +

0.5% arabinose and incubated in a 37°C shaker for 30 minutes. 500 uL of pTn5-v11 phage lysate was then added and the culture was incubated in a 37°C shaker for 20 minutes. The culture was then incubated statically at 37°C for 20 minutes. The phage infection reaction was then quenched with 500 uL of SOC + 200 mM sodium citrate and the culture was incubated in a 37°C shaker for 40 minutes. Cells were then pelleted by centrifugation and resuspended in 10 mL LB + gentamicin + hygromycin (50 ug/mL) + 0.5% arabinose and incubated in a 37°C shaker for 3 hours. Cells were then pelleted by centrifugation and resuspended for plating; all cells were plated on LB/agar plates supplemented with gentamicin and hygromycin, in full and as a dilution series.

SLIM pTn5 delivery/transposition, conjugation

Day 1: To begin each round, two colonies of SLIM target cells were picked from plates, either streak outs of glycerol stocks from previously picked colonies or colonies on plates from the end of the previous round of SLIM, and resuspended in 30 uL 10% LB. The resuspensions were used to independently inoculate 3 mL LB + gentamicin + 0.5% arabinose (to pre-induce cells for Tn5 transposition) cultures which were incubated in a 37°C shaker overnight. The donor strain, genRK24_piR + pTn5-v5 is grown in 3 mL LB + carbenicillin (100 ug/mL) chloramphenicol (15 ug/mL), kanamycin (50 ug/mL), hygromycin and 2% glucose.

Day 2: The next day, 500 uL from each of the SLIM target cell cultures was combined to form the conjugation recipient culture. This culture was pelleted by centrifugation and resuspended in 20 uL LB. 2.5 mL from the donor overnight culture was pelleted by centrifugation and resuspended in 100 uL LB. 10 uL from each of the donor and recipient resuspension was combined, mixed by pipetting and spotted in single 12 uL spots on LB/agar plates. When independent conjugation reactions were processed in parallel, donor and recipient mixtures were spotted on 1 mL LB/agar in a single well (per reaction) of a 24-well plate. The conjugation reaction was incubated, inverted, at 30°C for 2 hours. Spots were then washed using

1 mL and resuspended in 3 mL total LB+ gentamicin (30 ug/mL) + hygromycin and recovered in a 37°C shaker for 2 hours. Following recovery, cells were pelleted by centrifugation and plated in full and as a dilution series on LB/agar plates supplemented with gentamicin, hygromycin and 0.5% arabinose.

SLIM pTn5 delivery/transposition, electroporation

Day 1: To begin each round, two colonies of SLIM target cells were picked from plates, either streak-outs of glycerol stocks from previously picked colonies or colonies on plates from the end of the previous round of SLIM, and resuspended in 30 uL 10% LB. The resuspensions were used to independently inoculate 3 mL LB + gentamicin cultures which were incubated in a 37°C shaker overnight.

Day 2: The next day, 150 uL of each overnight culture was combined and used to inoculate a 15 mL LB + gentamicin culture in a 50 mL conical tube. The culture was grown to an OD between 0.45 and 0.6 prior to washing. Cells were then pelleted by centrifugation at 4000 rcf for 10 minutes at 4°C. The pellet was then resuspended in 1 mL of ice-cold H2O. Cells were washed using this procedure two additional times. Following the final wash, cells were pelleted by centrifugation and resuspended in 100 uL ice cold H2O. Prior to electroporation, 3.5 uL of pTn5v2 plasmid DNA was added to the cell resuspension and mixed by inverting. The cells/DNA combination was then added to a 2 mm electroporation cuvette and incubated on ice for 5 minutes. Cells were electroporated using an Eporator (Eppendorf) at 2.5 kV and immediately placed back on ice. Following electroporation, cells were recovered in 1 mL LB + 0.5% glucose for 1 hour in a 37°C shaker then transferred to a 10 mL LB + gentamicin and hygromycin culture and incubated overnight in a 37°C shaker.

Day 3: The next day, the overnight culture was used to inoculate a new 10 mL LB + gentamicin and hygromycin culture to an OD_{600} of 0.05. The culture was incubated in a 37°C shaker for 2 hours. To induce transposition, arabinose was

added to a final concentration of 0.5% and the culture was incubated in a 37°C shaker for an additional 2 hours. To select for transposition events and against maintenance of pTn5-v2 using sacB-mediated counterselection, sucrose was added to the culture to a final concentration of 7.5% and the culture was incubated in a 30°C shaker for 4 hours. Following incubation, cells were pelleted by centrifugation and plated in full and as a dilution series on LB/agar plates supplemented with gentamicin and hygromycin.

SLIM Cas3-mediated DNA removal

Day 1: Following delivery/transposition, four GFP+ colonies were selected and resuspended in 30 uL 10% LB for screening and cas3 induction.

For screening following electroporation-mediated delivery and transposition, 2 uL of each colony resuspension was spotted on two LB/agar plates; one supplemented with gentamicin and hygromycin (control), another supplemented with gentamicin, hygromycin and chloramphenicol to screen for loss of pTn5-v2.

For screening following phage transduction or conjugation-mediated delivery and transposition, 2 uL of each colony resuspension was spotted on three LB/agar plates; one supplemented with gentamicin and hygromycin (control), another supplemented with carbenicillin and chloramphenicol (donor screen) and a final plate supplemented with gentamicin, hygromycin and kanamycin to screen for loss of pTn5-v11 or pTn5-v5, respectively.

For cas3 induction, 15 uL of each colony resuspension was used to independently inoculate a 3 mL LB + gentamicin and hygromycin culture. Cells were grown overnight in a 37°C shaker.

Day 2: The next day, two of the four resuspended colonies were selected, based on screening results, to move forward to cas3 induction. For cas3 induction, 1.5 mL of each independent overnight culture was combined, pelleted by centrifugation and

resuspended in 3 mL LB + gentamicin + 0.3% rhamnose. The culture was then incubated in a 37°C shaker for 5.5 hours. Following incubation, cells were pelleted by centrifugation and plated in full and as a dilution series on LB/agar plates supplemented with gentamicin and 5 mM 4-chlorophenylalanine (Sigma-Aldrich).

Cas3-mediated DNA removal, Pseudomonas putida variation

Cas3-mediated DNA removal in *Pseudomonas putida* was conducted in a similar manner as described above with one exception; following cas3 induction and incubation, cells were pelleted by centrifugation and plated in full and as a dilution series on LB/agar plates supplemented with gentamicin only, without 4-chlorophenylalanine. The next day, colonies exhibiting cas3 activity through loss of GFP expression were selected for further screening and analysis.

pTn5 packaging in phage particles

Prior to delivery, pTn5 was packaged into phage particles *in vivo*, as previously described[94], in bulk for use in multiple rounds of SLIM. Briefly, pTn5-v11 was transformed into the P1 lysogenic strain p1kc_piR-sC. p1kc_piR-sC containing pTn5-v11 was grown in 3mL LB + chloramphenicol, hygromycin and kanamycin in a shaking incubator overnight at 37°C. The next day, a new 10 mL culture with ePLM + chloramphenicol, hygromycin and kanamycin was inoculated using the aforementioned overnight culture to an OD_{600} of 0.05 and incubated in a shaking incubator at 37°C. Upon reaching an OD between 0.8 and 1.0 the P1 lysogenic cycle was induced with 20 uM anhydrotetracycline. Following induction, the culture was incubated for an additional 2 hours in a shaking incubator at 37°C. After two hours, chloroform was added to the culture at a ratio of 1:40 chloroform to culture. The chloroform/culture mixture was incubated on ice for 5 minutes and mixed occasionally by pipetting. Following incubation, the culture was pelleted by centrifugation for 10 minutes at 3000 rcf at 4°C. The supernatant, phage lysate, was collected and stored at 4°C; the pellet was discarded.

Target cassette landing site identification

Delivery and transposition assays were coupled with a modified transposon sequencing (TnSeq) protocol to identify genomic landing sites for the target cassette. Delivery, using phage transduction, conjugation or electroporation, and transposition assays were performed as described above. Approximately 200 colonies were selected from post-transposition plates, per delivery method, pooled for analysis. gDNA was extracted according to the manufacturer's instructions (QIAGEN™ DNeasy Blood & Tissue Kit cat #69504). Concentrations of gDNA were quantified with the Invitrogen Qubit™ 4 Fluorometer. DNA libraries were prepared for modified TnSeq as previously described[54]. Pooled library denaturation and flow cell loading was performed per manufacturer instructions. Prior to sequencing, a custom sequencing primer, designed to bind to the target cassette, was spiked into the reagent kit at a final concentration of 0.5 μM to enable targeted sequencing of enriched transposon sequences within the libraries. Sequencing was performed on an Illumina MiSeq using MiSeq reagent kit v3 (600-cycles) with 300 bp paired-end reads and automated demultiplexing and adapter trimming.

Output reads were first aligned to a target cassette reference using BWA short-read sequence aligner[95]. All reads which aligned to the target cassette reference were collected using reformat.sh[96] and then aligned to a reference *E. coli* DH10B genome (retrieved from NCBI RefSeq, assembly GCF_000019425.1) using BWA. Locations of aligned reads were annotated then plotted using the pyCirclize python package[97].

Cas3-mediated DNA removal by non-essential span

Experiments to assess efficacy of Cas3-mediated DNA removal were conducted using four independent strains. Progenitor strains containing pKW20 were modified through precise integration of the original (target sequence, hygR and GFP) or modified (target sequence, pheS^{mut}, hygR and GFP) target cassette using a

Lambda-Red recombination system encoded by pKW20. After sequence and genomic location of the target cassette was confirmed for each strain, pKW20 was cured and pCas3 was transformed. For Cas3-mediated DNA removal, modified progenitor strains containing pCas3 and target cassettes were grown overnight in 3 mL LB + hygromycin and gentamicin. The next day, overnight cultures were pelleted by centrifugation for 6 minutes at 4000g, the supernatant was discarded and pellets were resuspended in fresh LB + gentamicin and 0.3% rhamnose for Cas3 induction. Cultures were incubated in a 37°C shaker for 5.5 hours, then plated on LB/agar plates supplemented with just gentamicin (original target cassette) or gentamicin + 4-chlorophenylalanine (modified target cassette).

Following Cas3-mediated DNA removal, the number of GFP negative colonies were counted and divided by the number of total colonies to get the GFP- ratio for each strain/cassette type. To assess potential genomic DNA removal, four individual GFP- colonies were picked per strain. Genomic DNA was extracted as previously described from each colony independently and extracted gDNA libraries were prepared for whole genome sequencing (WGS) according to manufacturer's instructions (Illumina). gDNA libraries were sequenced using the Illumina MiSeq according to manufacturer's instructions. 300bp reads output from MiSeq sequencing were aligned to a reference *E. coli* DH10B genome and gaps in alignment were identified as DNA removed by Cas3.

Sample and library preparation for whole genome sequencing (WGS)

WGS of all WT and genome-reduced strains was performed as previously described[54]. Briefly, bacterial cultures were grown in LB media supplemented with the appropriate antibiotics, until confluent. Cultures were pelleted by centrifugation and genomic DNA (gDNA) was extracted from the pellets using DNeasy Blood & Tissue Kit (QIAGEN™ cat # 69504) per manufacturer instructions. Concentrations of gDNA were quantified with the Invitrogen Qubit™ 4 Fluorometer using the 1x dsDNA High Sensitivity (HS) assay kit (Thermo Fisher Scientific cat # Q33231).

For library preparation, Nextera™ DNA Flex Library Prep Kit (Illumina) was used per manufacturer instructions to tag, barcode, amplify and add index primers (i5 and i7) to the library. 200 ng of starting gDNA was used per barcoded library. Prepared libraries were quantified with the Qubit™ 4 Fluorometer before pooling 15 ng per barcoded library. Pooled library denaturation and flow cell loading was performed per manufacturer instructions. Sequencing was performed on an Illumina MiSeq using MiSeq reagent kit v3 (600-cycles) with 300 bp paired-end reads and automated demultiplexing and adapter trimming.

WGS data analysis

Sequencing reads from whole-genome samples were aligned to expected reference genomes using the BWA short-read sequence aligner[95]. Alignment files were further processed with Samtools[98] and deepTools[99] was used to estimate RPGC throughout the expected reference genome in 10-kb bins. Regions of low to zero coverage were analyzed in higher resolution using Interactive Genome Viewer[100]. Exact endpoints of removed DNA segments were identified using IGV, to single nucleotide resolution.

Coverage Plots

Annotated regions of removed DNA, for each genome-reduced strain, were used to generate coverage plots. Each nucleotide, per strain, was assigned a 0 if removed or a 1 retained. Resulting values were plotted for each genome-reduced strain independently and as the average across all genome-reduced strains.

Deletion permissibility assessments

We assessed deletion permissibility based on two characteristics, function and genomic location, independently.

Deletion permissibility assessment, by function

To assess deletion permissibility based on function we grouped genes into functional groups based on COG group designation. Following COG group segregation, we computed the percentage of genes deleted in each group on a per line basis. For the initial assessment, we performed qualitative comparisons of the percentage of deleted genes across genome-reduced strains per group. Next, we computed the median percentage of deleted genes for each group.

Deletion permissibility assessment, by genomic location

To assess deletion permissibility based on genomic location we needed to generate deletion probabilities for each individual gene, conduct deletion simulations and then assess the simulations based on two grouping strategies.

To generate deletion probabilities we computed the average coverage per nucleotide based on deletions generated in the 10 member library through 21 rounds of SLIM. Round 0 deletions and deletions in one half of the tandem repeat section were excluded from these analyses. For any nucleotide in the genome, if the nucleotide had not been deleted for a given genome-reduced strain, it was given a value of 1. If the nucleotide had been deleted, it was given a value of 0. The average value for each nucleotide was then computed across all genome-reduced strains. Next, the genome was segmented into bins according to the desired number of bins. The average coverage per bin was computed and converted to average deletion probability by subtracting average coverage from 1. Each gene in any particular bin was then assigned a deletion probability equivalent to the average deletion probability of the bin.

Using computed deletion probabilities we simulated genome-wide deletions on a per gene basis. For each gene we generated a random number between 0 and 1. If the randomly generated number was less than the deletion probability, that gene

was indicated as deleted in the simulation. Otherwise, the gene was indicated as retained. Deletion predictions were accumulated across 10 simulations for each prediction assessment and the result of each simulation was recorded independently per gene for each simulation.

For randomized deletion probability predictions, computed deletion probabilities were scrambled and randomly assigned to individual genes. Randomized deletion probability predictions were carried out in the same manner as described for computed deletion probability predictions.

Following deletion simulations we randomly assigned each gene to one of 25 groups. The percentage of deleted genes was computed, per group, based on observed deletions for each genome-reduced strain and based on simulation outputs utilizing each of computed and randomized deletion probability predictions.

Deletion permissibility assessment based on genomic location and COG grouping was conducted as described for random grouping, except genes were assigned to groups according to COG group indications.

For each group the set of the percentage of deleted genes based on observed deletions was compared to each of the percentage of deleted genes sets from simulated deletions based on computed deletion probability predictions and randomized deletion probability predictions. The coefficient of determination (R^2) was computed using the r2_score function from scikit-learn, for the observed set and each predicted set independently, on a per-group basis.

Sample preparation for transcriptomic analysis

E. coli were grown overnight in 10 mL LB + gentamicin in a shaking incubator at 37°C to saturation. The next day, 1 mL of overnight culture was collected, pelleted by centrifugation and resuspended in 100 uL of ice cold TE buffer. 2 uL Readylyse lysozyme (Biosearch Technologies) was added to each sample. Samples were then

incubated on a shaking heat block for 10 minutes at 25°C and 500 RPM. Following incubation each sample was mixed with 300 uL lysis buffer and RNA extractions were performed according to manufacturer's instructions (Monarch Total RNA Miniprep Kit, New England Biolabs). Extracted RNA was quantified using Invitrogen Qubit™ 4 Fluorometer using the RNA High Sensitivity (HS) assay kit (Thermo Fisher Scientific cat # Q32852). rRNA depletion was performed and sequencing libraries from extracted RNA samples were prepped and sequenced on an Illumina NextSeq2000 with 50 bp paired-end reads by the Millard and Muriel Jacobs Genetics and Genomics Laboratory at the California Institute of Technology. RNA-Seq experiments were performed with three biological replicates per strain.

Transcriptomic data analysis

RNA sequencing output reads were aligned to a reference transcriptome (retrieved from NCBI RefSeq, assembly GCF_000019425.1, and further processed in-house) and transcript abundances were quantified using Kallisto[60]. Transcript abundances, in the form of transcripts per million (TPM), output from Kallisto analyzed further using an in-house Python script. Briefly, abundances were median normalized across biological replicates on a per genome-reduced strain combined with WT basis. T-statistics and corresponding p-values were computed using SciPy v1.16.0. Thresholds for differential expression were set at p-value < 0.05 and a fold-change ≥ 1 for upregulation and p-value < 0.05 and a fold-change < -1 for downregulation. The results of differential expression analyses, on a per genome-reduced strain basis, were plotted on volcano plots.

Data from single transcript differential analyses were used to evaluate differential expression of select pathways. Pathway gene lists were curated by cross-referencing lists from KEGG[64], amiGO[66] and string-db[67]. Pathway scores were computed as the sum of $\log_2$FC values for each transcript product of genes in the pathway of interest. Scores were computed for each genome-reduced strain

individually. Pathway scores were expressed on heatmaps as temperatures. Parallel analyses were conducted for all pathways relating to the status of deleted genes, both with and without deletion penalties. For analyses conducted with deletion penalties, a $\log_2$FC value of -10 was awarded to genes that had been deleted during the genome reduction process. Assignment of deletion penalties was done individually for each genome-reduced strain. Analyses conducted without deletion penalties simply removed deleted genes from the computation of pathway expression scores. Pathway gene list adjustments were done independently for each genome-reduced strain.

Sample preparation for proteomic analysis

Three biological replicates per strain were grown overnight in 10 mL LB + gentamicin in a shaking incubator at 37°C to saturation. The next day, overnight cultures were collected and pelleted by centrifugation. Cell pellets were resuspended in 5% sodium dodecyl sulfate (Sigma-Aldrich) in 50 mM HEPES, and were homogenized using BeatBox (Preomics) for 10 min under 'High' settings. Protein concentration was measured using Pierce BCA protein assay kit (Pierce), and 100 µg of protein was used for further sample preparation. The samples were reduced using 5 mM tris(2-carboxyethyl)phosphine (Sigma-Aldrich) under 55°C for 10min, and then alkylated with chloroacetamide (Sigma-Aldrich) under room temperature for 15min. The samples were further acidified to a final concentration of 2.5% phosphoric acid, and 25 µl of sample was combined with 165 µl of 90% methanol with 10% of 1 M triethylammonium bicarbonate (TEAB, Thermo Scientific) according to previous protocol (PMID: 30114372). The samples were then loaded onto S-trap (Protifi) devices, and the S-trap devices were washed using the same buffer for loading (90% methanol with 10% of 1 M TEAB) for 3 times. For each loading or washing step, the S-trap devices were centrifuged at 4000 g for 30 seconds to remove the elute. After washing steps, 20 µl of 100 mM TEAB containing 10 µg of TPCK-trypsin (Thermo Scientific) was added into each sample,

and the digestion was allowed overnight. The digested peptides were eluted using 40 µl of 50 mM TEAB in water, 0.2% formic acid in water, and 50% acetonitrile in water sequentially with 4000 g centrifugation for 1 min for each elution. The elutes for 3 steps were pooled together and dried using a refrigerated CentiVap concentrator (Labconco). The dried samples were stored in -20°C before resuspended in mobile phase A (2% acetonitrile, 0.2% formic acid, and 97.8% water) for LC-MS/MS analysis.

LC-MS/MS for proteomic analysis

For proteomic samples, LC-MS/MS experiments were performed by loading 500 ng sample onto an EASY-nLC 1200 (ThermoFisher Scientific, San Jose, CA) connected to an Q Exactive HF Quadrupole – Orbitrap Hybrid mass spectrometer (Thermo Fisher Scientific, San Jose, CA). Peptides were separated on an Aurora Ultimate XT UHPLC column (25 cm × 75 µm, 1.6 µm C18, AUR3-25075C18-XT, Ion Opticks) with a flow rate of 0.35 µL/min and for a total duration of 131 min. The gradient was composed of 3% Solvent B for 1 min, 3–19% B for 72 min, 19–29% B for 28 min, 29–41% B for 20 min, 41–95% B for 3 min, and 95–98% B for 7 min. Solvent A consists of 97.8% H2O, 2% ACN, and 0.2% formic acid, and solvent B of 19.8% H2O, 80% ACN, and 0.2% formic acid. MS1 scans were acquired with a range of 375–1500 m/z in the Orbitrap at 60 k resolution. The maximum injection time was 15 ms, and the AGC target was 3 × 106. MS2 scans were acquired at 30 k resolution with a first scan mass as 100 Da. The maximum injection time was 45 ms, and the AGC target was 3 × 106. The isolation window was 1.2 m/z, collision energy was 28 NCE, and loop count was 12. Other global settings were set to the following: ion source type, NSI; spray voltage, 2000 V; ion transfer tube temperature, 300°C. Method modification and data collection were performed using Xcalibur software (Thermo Scientific).

Proteomics data analysis

Proteomic analysis was performed using Proteome Discoverer 2.5 (PD 2.5, Thermo Scientific) software, and SequestHT with Percolator validation. The raw data was searched against *Escherichia coli* proteome (retrieved from NCBI RefSeq, assembly GCF_000019425.1). Percolator FDRs were set at 0.001 (strict) and 0.005 (relaxed). Peptide FDRs were set at 0.001 (strict) and 0.005 (relaxed), with medium confidence and a minimum peptide length of 6. Carbamidomethyl (C) was set as a static modification; oxidation (M) was set as a dynamic modification; acetyl (protein N-term), Met-loss (Protein N-term M) and Met-loss + acetyl (Protein N-term M) were set as dynamic N-Terminal modifications.

Further analysis was performed using an in-house python script to assess the differential expression of individual protein products. Briefly, all quantitation from match-between-run analysis was eliminated to prevent false discovery. The data were further normalized to each sample's median value for differential expression analysis. T-statistics and corresponding p-values were computed using SciPy v1.16.0. Thresholds for differential expression were set at p-value < 0.05 and a fold-change ≥ 1 for upregulation and p-value < 0.05 and a fold-change < -1 for downregulation. The results of differential expression analyses, on a per genome-reduced strain basis, were plotted on volcano plots.

Data from single protein differential analyses were used to evaluate differential expression of select pathways. Pathway gene lists were curated by cross-referencing lists from KEGG[64], amiGO[66] and string-db[67]. Pathway scores were computed as the sum of $\log_2$FC values for each protein product of genes in the pathway of interest. Scores were computed for each genome-reduced strain individually. Pathway scores were expressed on heatmaps as temperatures. Parallel analyses were conducted for all pathways relating to the status of deleted genes, both with and without deletion penalties. For analyses conducted with deletion penalties, a $\log_2$FC value of -10 was awarded to genes that had been deleted during

the genome reduction process. Assignment of deletion penalties was done individually for each genome-reduced strain. Analyses conducted without deletion penalties simply removed deleted genes from the computation of pathway expression scores. Pathway gene list adjustments were done independently for each genome-reduced strain.

RNA transcript and protein accounting

RNA transcript counts and protein counts were acquired and normalized as previously described. Following normalization, counts for each transcript and protein among biological replicates from each strain were averaged on a per strain basis. For each strain, the sum of counts per COG group was computed and then normalized according to COG group size. Sums across COG groups, per strain, were computed. Individual COG group ratios were computed by dividing the group totals for transcripts or proteins by the total transcripts or proteins counted across all groups for each individual strain.

Plate Reader Growth Assay

The progenitor and all ten genome-reduced strains, each containing pCas3, were grown overnight from glycerol stocks in 3 mL LB + gentamicin. The next day 220 uL LB + gentamicin was added to 24 individual wells in a round-bottom 96 well plate. Each well was then inoculated with one of the eleven overnight cultures to a density of OD_{600} 0.05 at a rate of two wells per strain for a total of 22 wells. The additional two wells were used in downstream analyses, to subtract baseline OD_{600} values generated by LB alone. The plate was incubated in a Tecan Spark® Multimode Microplate Reader for 16 hours at a constant temperature of 37°C. OD_{600} measurements were taken every 15 minutes with shaking incubation implemented in between measurements with an amplitude of 2.5 mm and a double-orbital path.

Large Volume Growth Assays

Strains were grown overnight from glycerol stocks in 3 mL LB + requisite antibiotics. For growth experiments comparing WT to Progenitor, L7 and L8 strains containing pCas3, WT (DH10B) was grown in LB + streptomycin (100 ug/mL) and Progenitor, L7 and L8, all containing pCas3, were grown in gentamicin. The next day, overnight cultures were used to inoculate new, 50 mL cultures in 500 mL flasks containing either LB, LB + 2% glucose (LBg) or Terrific Broth (TB) and requisite antibiotics. 50 mL cultures were grown in a shaking incubator at 37°C for 24 hours. Every half hour for the initial 10 hours of growth, 500 uL samples were taken from each culture to measure OD_{600}. The final OD_{600} measurement was taken for each culture after 24 hours.

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

APPENDIX A

TABLE OF DELETED GENOMIC SEGMENTS

Strain	**Coordinates**	**Size**
All Strains (rd0)	248,478..309,298	60,821
L1	514,313..627,568	113,256
L1	735,072..739,721	4650
L1	910,203..970,768	60,566
L1	1,446,335..1,472,745	26,411
L1	1,478,719..1,484,546	5828
L1	1,522,966..1,531,578	8613
L1	1,532,070..1,533,576	1507
L1	1,577,553..1,622,608	45,056
L1	3,095,393..3,109,178	13,786
L1	3,202,149..3,214,357	12,209
L1	3,355,526..3,360,477	4952
L1	3,589,174..3,607,682	18,509
L1	3,755,464..3,770,393	14,930
L1	3,838,007..3,874,032	36,026
L1	4,168,448..4,186,994	18,547
L1	4,389,294..4,405,915	16,622
L1	4,591,361..4,644,364	53,004
L2	330,543..361,830	31,288
L2	400,888..402,245	1358
L2	514,313..627,568	113,256
L2	832,845..845,530	12,686
L2	855,172..863,066	7895
L2	1,348,860..1,397,857	48,998
L2	1,589,819..1,599,674	9856
L2	1,755,492..1,772,031	16,540
L2	1,830,433..1,831,124	692

L2	2,203,830..2,206,699	2870
L2	2,557,511..2,567,725	10,215
L2	2,850,764..2,896,946	46,183
L2	3,119,399..3,125,312	5914
L2	3,635,303..3,641,667	6365
L2	3,706,471..3,712,036	5566
L2	3,842,140..3,852,356	10,217
L2	3,855,898..3,874,495	18,598
L2	4,599,279..4,640,717	41,439
L3	514,313..627,568	113,256
L3	1,218,350..1,218,424	75
L3	1,251,185..1,280,884	29,700
L3	1,349,013..1,397,832	48,820
L3	1,445,493..1,448,188	2696
L3	1,483,658..1,522,728	39,071
L3	1,540,968..1,678,465	137,498
L3	1,712,393..1,731,497	19,105
L3	1,848,997..1,866,880	17,884
L3	2,889,302..2,897,480	8179
L3	3,195,373..3,200,817	5445
L3	3,202,149..3,214,162	12,014
L3	3,457,967..3,462,671	4705
L3	3,737,134..3,749,152	12,019
L3	3,791,428..3,819,137	27,710
L3	3,855,935..3,875,015	19,081
L3	4,028,966..4,035,978	7013
L3	4,601,195..4,619,652	18,458
L4	331,522..336,987	5466
L4	514,313..627,568	113,256
L4	755,983..756,488	506
L4	786,639..804,587	17,949

L4	902,435..974,939	72,505
L4	1,546,851..1,561,198	14,348
L4	1,590,166..1,622,365	32,200
L4	1,644,135..1,677,195	33,061
L4	2,546,095..2,549,761	3667
L4	2,848,680..2,854,427	5748
L4	3,192,736..3,200,814	8079
L4	3,202,149..3,213,275	11,127
L4	3,585,876..3,604,347	18,472
L4	3,719,495..3,742,959	23,465
L4	3,837,936..3,871,180	33,245
L4	4,159,500..4,167,679	8180
L4	4,590,283..4,641,938	51,656
L5	1,682,394..1,698,853	16,460
L5	1,785,485..1,801,299	15,815
L5	2,847,422..2,848,646	1225
L5	3,722,691..3,817,577	94,887
L5	3,837,759..3,878,532	40,774
L5	4,157,441..4,167,676	10,236
L5	4,301,284..4,322,829	21,546
L6	514,313..627,568	113,256
L6	755,984..756,499	516
L6	918,152..964,837	46,686
L6	1,325,192..1,329,514	4323
L6	1,508,605..1,551,819	43,215
L6	1,644,135..1,677,195	33,061
L6	1,699,220..1,712,420	13,201
L6	1,758,419..1,764,903	6485
L6	2,247,288..2,269,694	22,407
L6	2,546,095..2,549,761	3667
L6	2,848,680..2,855,825	7146

L6	3,166,482..3,171,156	4675
L6	3,214,607..3,241,893	27,287
L6	3,577,527..3,581,395	3869
L6	3,585,363..3,603,738	18,376
L6	3,794,166..3,817,581	23,416
L6	3,835,502..3,853,420	17,919
L6	3,995,746..4,002,291	6546
L6	4,382,247..4,388,525	6279
L7	447,589..463,418	15,830
L7	585,548..624,911	39,364
L7	1,348,736..1,397,676	48,941
L7	1,414,040..1,416,922	2883
L7	1,451,456..1,605,432	153,977
L7	1,906,701..1,907,631	931
L7	2,082,611..2,107,573	24,963
L7	2,343,565..2,346,754	3190
L7	2,412,883..2,418,078	5196
L7	3,168,874..3,170,543	1670
L7	3,202,149..3,223,630	21,482
L7	3,376,480..3,402,553	26,074
L7	3,647,821..3,657,912	10,092
L7	3,940,322..3,943,595	3274
L7	4,028,111..4,034,434	6324
L7	4,107,426..4,114,268	6843
L7	4,389,295..4,412,573	23,279
L8	444,330..475,983	31,654
L8	514,313..627,568	113,256
L8	921,252..974,006	52,755
L8	1,348,644..1,397,893	49,250
L8	1,689,072..1,729,845	40,774
L8	1,906,701..1,907,632	932

L8	2,847,791..2,896,946	49,156
L8	3,200,768..3,214,415	13,648
L8	3,572,879..3,588,504	15,626
L8	3,635,288..3,641,652	6365
L8	3,642,728..3,658,032	15,305
L8	3,722,709..3,817,581	94,873
L8	4,028,111..4,034,434	6324
L8	4,168,449..4,191,832	23,384
L8	4,388,776..4,392,174	3399
L8	4,636,624..4,644,362	7739
L9	456,607..490,357	33,751
L9	514,313..627,568	113,256
L9	647,631..651,623	3993
L9	669,421..689,982	20,562
L9	1,324,984..1,330,658	5675
L9	1,348,769..1,398,180	49,412
L9	1,447,706..1,503,067	55,362
L9	1,528,191..1,559,116	30,926
L9	1,594,464..1,610,270	15,807
L9	1,726,152..1,750,416	24,265
L9	2,309,280..2,332,993	23,714
L9	2,560,066..2,589,699	29,634
L9	2,755,676..2,783,803	28,128
L9	3,740,288..3,781,962	41,675
L9	3,898,535..3,901,288	2754
L9	4,167,872..4,187,975	20,104
L9	4,374,147..4,389,130	14,984
L9	4,637,459..4,644,363	6905
L10	514,313..627,568	113,256
L10	911,662..965,084	53,423
L10	1,503,980..1,674,056	170,077

L10	1,680,231..1,687,263	7033
L10	1,713,176..1,730,196	17,021
L10	1,737,524..1,740,539	3016
L10	1,780,779..1,791,521	10,743
L10	1,808,046..1,812,467	4422
L10	2,557,511..2,567,725	10,215
L10	3,284,544..3,289,336	4793
L10	3,457,879..3,462,671	4793
L10	3,632,495..3,642,271	9777
L10	3,731,020..3,781,948	50,929
L10	3,945,724..3,970,479	24,756
L10	4,156,829..4,167,679	10,851
L10	4,317,579..4,344,277	26,699
L10	4,389,294..4,452,419	63,126
L10	4,637,076..4,641,938	4863
SL1	1,769,283..1,769,728	446
SL1	3,105,224..3,144,989	39,766
SL1	3,478,885..3,479,182	298
SL2	807,064..859,864	52,801
SL2	3,121,010..3,152,127	31,118
SL2	3,965,561..3,967,668	2108
SL3	500,478..534,804	34,327
SL3	2,178,448..2,211,256	32,809
SL3	3,083,001..3,128,297	45,297
PL1	869,849..880,781	10,933
PL1	3,820,981..4,068,420	247,440
PL1	4,790,351..4,824,366	34,016
PL2	4,531,459..4,619,852	88,394
PL3	2,285,330..2,293,514	8185
PL3	3,362,126..3,718,902	356,777

Appendix A: Table of all observed deleted genomic segments across all genome-reduced strains and all rounds of genome reduction using SLIM in *E. coli* (L1-10), *S. flexneri* (SL1-3) and *P. putida* (PL1-3).

APPENDIX B

TABLE OF DELETED ESSENTIAL GENES

Strain	Deleted Essential Genes (Any Study)	Deleted Essential Genes (Consensus)
L1	yagG', 'yncH', 'tdcF'	n/a
L2	yagG'	n/a
L3	yagG', 'yncH', 'ydfO', 'minD', 'ydiL', 'safA', 'racR', 'ydaS', 'iraM'	n/a
L4	yncH', 'safA', 'yagG'	n/a
L5	bcsB', 'yagG'	n/a
L6	safA', 'yagG'	n/a
L7	racR', 'ydaS', 'alsK', 'rsmI', 'yagG'	n/a
L8	yagG', 'ydfO', 'bcsB'	n/a
L9	relB', 'yagG', 'glyA', 'dicA'	n/a
L10	yncH', 'yagG', 'ydaS', 'alsK', 'safA', 'ydfO', 'racR'	n/a

Appendix B: Table of deleted essential genes. Essential gene indication was determined from three independent studies; KEIO[58], traDIS[49] and PEC[59]. Any gene that was deleted from the genome of any genome-reduced strain and was listed as essential by any one of the three studies is listed in 'Deleted Essential Genes (Any Study).' No consensus essential genes (determined to be essential by all three studies) were deleted during the genome reduction process.

BIBLIOGRAPHY

1. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. Chinwalla, A. T. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
3. Glover, N. The Banana Conjecture. *Dessimoz Lab* https://lab.dessimoz.org/blog/2020/12/08/human-banana-orthologs (2020).
4. Guenther, C. A., Tasic, B., Luo, L., Bedell, M. A. & Kingsley, D. M. A molecular basis for classic blond hair color in Europeans. *Nat. Genet.* **46**, 748–752 (2014).
5. Fischer, M. D. *et al.* Codon-Optimized RPGR Improves Stability and Efficacy of AAV8 Gene Therapy in Two Mouse Models of X-Linked Retinitis Pigmentosa. *Mol. Ther.* **25**, 1854–1865 (2017).
6. Lin, P., Jiang, J. & Wu, M. CRISPR base editor treats premature-aging syndrome. *Signal Transduct. Target. Ther.* **6**, 158 (2021).
7. Elangkovan, N. & Dickson, G. Gene Therapy for Duchenne Muscular Dystrophy. *J. Neuromuscul. Dis.* **8**, S303–S316.
8. Ribeil, J.-A. *et al.* Gene Therapy in a Patient with Sickle Cell Disease. *N. Engl. J. Med.* **376**, 848–855 (2017).
9. Ochman, H. & Jones, I. B. Evolutionary dynamics of full genome content in Escherichia coli. *EMBO J.* **19**, 6637–6643 (2000).

10. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).

11. Borodovich, T., Shkoporov, A. N., Ross, R. P. & Hill, C. Phage-mediated horizontal gene transfer and its implications for the human gut microbiome. *Gastroenterol. Rep.* **10**, goac012 (2022).

12. Desfarges, S. & Ciuffi, A. Viral Integration and Consequences on Host Gene Expression. *Viruses Essent. Agents Life* 147–175 (2012) doi:10.1007/978-94-007-4899-6_7.

13. Lally, P. *et al.* A Cryptic Prophage Transcription Factor Drives Phenotypic Changes via Host Gene Regulation. *bioRxiv* 2024.09.21.614188 (2024) doi:10.1101/2024.09.21.614188.

14. Siguier, P., Gourbeyre, E. & Chandler, M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol. Rev.* **38**, 865–891 (2014).

15. Koskiniemi, S., Sun, S., Berg, O. G. & Andersson, D. I. Selection-Driven Gene Loss in Bacteria. *PLoS Genet.* **8**, e1002787 (2012).

16. Pedersen, I. B., Helgesen, E., Flåtten, I., Fossum-Raunehaug, S. & Skarstad, K. SeqA structures behind Escherichia coli replication forks affect replication elongation and restart mechanisms. *Nucleic Acids Res.* **45**, 6471–6485 (2017).

17. Bzymek, M., Saveson, C. J., Feschenko, V. V. & Lovett, S. T. Slipped Misalignment Mechanisms of Deletion Formation: In Vivo Susceptibility to Nucleases. *J. Bacteriol.* **181**, 477–482 (1999).

18. Ireland, W. T. *et al.* Deciphering the regulatory genome of Escherichia coli, one hundred promoters at a time. *eLife* **9**, e55308 (2020).

19. Sharma, S. S., Blattner, F. R. & Harcum, S. W. Recombinant protein production in an *Escherichia coli* reduced genome strain. *Metab. Eng.* **9**, 133–141 (2007).

20. Lieder, S., Nikel, P. I., de Lorenzo, V. & Takors, R. Genome reduction boosts heterologous gene expression in Pseudomonas putida. *Microb. Cell Factories* **14**, 23 (2015).

21. Vollmann, D. J., Lernoud, L. & Nett, M. Genome Reduction Improves Recombinant Benzoxazole Production in Myxococcus xanthus. *ACS Synth. Biol.* **14**, 1756–1765 (2025).

22. Morimoto, T. *et al.* Enhanced Recombinant Protein Productivity by Genome Reduction in Bacillus subtilis. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* **15**, 73–81 (2008).

23. Kurasawa, H., Ohno, T., Arai, R. & Aizawa, Y. A guideline and challenges toward the minimization of bacterial and eukaryotic genomes. *Curr. Opin. Syst. Biol.* **24**, 127–134 (2020).

24. Hutchison, C. A. *et al.* Global transposon mutagenesis and a minimal Mycoplasma genome. *Science* **286**, 2165–2169 (1999).

25. Kobayashi, K. *et al.* Essential Bacillus subtilis genes. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 4678–4683 (2003).

26. Hutchison, C. A. *et al.* Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016).

27. Zhou, J., Wu, R., Xue, X. & Qin, Z. CasHRA (Cas9-facilitated Homologous Recombination Assembly) method of constructing megabase-sized DNA. *Nucleic Acids Res.* **44**, e124 (2016).

28. Xu, X. *et al.* Trimming the genomic fat: minimising and re-functionalising genomes using synthetic biology. *Nat. Commun.* **14**, 1984 (2023).

29. Pósfai, G. *et al.* Emergent Properties of Reduced-Genome Escherichia coli. *Science* **312**, 1044–1046 (2006).

30. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 6640–6645 (2000).

31. Karcagi, I. *et al.* Indispensability of Horizontally Transferred Genes and Its Impact on Bacterial Genome Streamlining. *Mol. Biol. Evol.* **33**, 1257–1269 (2016).

32. Mizoguchi, H., Sawano, Y., Kato, J. & Mori, H. Superpositioning of Deletions Promotes Growth of Escherichia coli with a Reduced Genome. *DNA Res.* **15**, 277–284 (2008).

33. Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y. & Ishikawa, H. Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. *Nature* **407**, 81–86 (2000).

34. Hirokawa, Y. *et al.* Genetic manipulations restored the growth fitness of reduced-genome *Escherichia coli*. *J. Biosci. Bioeng.* **116**, 52–58 (2013).

35. Reuß, D. R. *et al.* Large-scale reduction of the Bacillus subtilis genome: consequences for the transcriptional network, resource allocation, and metabolism. *Genome Res.* **27**, 289–299 (2017).

36. Komatsu, M., Uchiyama, T., Ōmura, S., Cane, D. E. & Ikeda, H. Genome-minimized Streptomyces host for the heterologous expression of secondary metabolism. *Proc. Natl. Acad. Sci.* **107**, 2646–2651 (2010).

37. Yu, B. J. *et al.* Minimization of the Escherichia coli genome using a Tn5-targeted Cre/loxP excision system. *Nat. Biotechnol.* **20**, 1018–1023 (2002).

38. Cui, L. & Bikard, D. Consequences of Cas9 cleavage in the chromosome of Escherichia coli. *Nucleic Acids Res.* **44**, 4243–4251 (2016).

39. Chayot, R., Montagne, B., Mazel, D. & Ricchetti, M. An end-joining repair mechanism in Escherichia coli. *Proc. Natl. Acad. Sci.* **107**, 2141–2146 (2010).

40. Vernyik, V. *et al.* Exploring the fitness benefits of genome reduction in Escherichia coli by a selection-driven approach. *Sci. Rep.* **10**, 7345 (2020).

41. Ma, S., Su, T., Liu, J., Lu, X. & Qi, Q. Reduction of the Bacterial Genome by Transposon-Mediated Random Deletion. *ACS Synth. Biol.* **11**, 668–677 (2022).

42. Bowater, R. & Doherty, A. J. Making Ends Meet: Repairing Breaks in Bacterial DNA by Non-Homologous End-Joining. *PLOS Genet.* **2**, e8 (2006).

43. Vercoe, R. B. *et al.* Cytotoxic Chromosomal Targeting by CRISPR/Cas Systems Can Reshape Bacterial Genomes and Expel or Remodel Pathogenicity Islands. *PLoS Genet.* **9**, e1003454 (2013).

44. Rancati, G., Moffat, J., Typas, A. & Pavelka, N. Emerging and evolving concepts in gene essentiality. *Nat. Rev. Genet.* **19**, 34–49 (2018).

45. Sauder, A. B. & Kendall, M. M. After the Fact(or): Posttranscriptional Gene Regulation in Enterohemorrhagic Escherichia coli O157:H7. *J. Bacteriol.* **200**, e00228-18 (2018).

46. Mateus, A. *et al.* Transcriptional and Post-Transcriptional Polar Effects in Bacterial Gene Deletion Libraries. *mSystems* **6**, e00813-21.

47. Pal, A., Iyer, M. S., Srinivasan, S., Narain Seshasayee, A. S. & Venkatesh, K. V. Global pleiotropic effects in adaptively evolved Escherichia coli lacking CRP reveal molecular mechanisms that define the growth physiology. *Open Biol.* **12**, 210206.

48. Goryshin, I. Y. & Reznikoff, W. S. Tn5 in Vitro Transposition. *J. Biol. Chem.* **273**, 7367–7374 (1998).

49. Goodall, E. C. A. *et al.* The Essential Genome of Escherichia coli K-12. *mBio* **9**, 10.1128/mbio.02096-17 (2018).

50. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).

51. Csörgő, B. *et al.* A compact Cascade–Cas3 system for targeted genome engineering. *Nat. Methods* **17**, 1183–1190 (2020).

52. Campbell, B. C. *et al.* mGreenLantern: a bright monomeric fluorescent protein with rapid expression and cell filling properties for neuronal imaging. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 30710–30721 (2020).

53. Stalker, D. M., Kolter, R. & Helinski, D. R. Plasmid R6K DNA replication: I. Complete nucleotide sequence of an autonomously replicating segment. *J. Mol. Biol.* **161**, 33–43 (1982).

54. Sanfiorenzo, C. *et al.* Creation of genomic chimeras via megabase scale genome expansion. *Nature* In revision (2025).

55. Miyazaki, K. Molecular Engineering of a PheS Counterselection Marker for Improved Operating Efficiency in Escherichia Coli. *BioTechniques* **58**, 86–88 (2015).

56. Durfee, T. *et al.* The Complete Genome Sequence of Escherichia coli DH10B: Insights into the Biology of a Laboratory Workhorse. *J. Bacteriol.* **190**, 2597–2606 (2008).

57. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).

58. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).

59. Hashimoto, M. *et al.* Cell size and nucleoid organization of engineered Escherichia coli cells with a reduced genome. *Mol. Microbiol.* **55**, 137–149 (2005).

60. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

61. Chib, S. & Mahadevan, S. Involvement of the global regulator H-NS in the survival of Escherichia coli in stationary phase. *J. Bacteriol.* **194**, 5285–5293 (2012).

62. Nair, S. & Finkel, S. E. Dps protects cells against multiple stresses during stationary phase. *J. Bacteriol.* **186**, 4192–4198 (2004).

63. Khil, P. P. & Camerini-Otero, R. D. Over 1000 genes are involved in the DNA damage response of Escherichia coli. *Mol. Microbiol.* **44**, 89–105 (2002).

64. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

65. Moore, L. R. *et al.* Revisiting the y-ome of Escherichia coli. *Nucleic Acids Res.* **52**, 12201–12207 (2024).

66. Carbon, S. *et al.* AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**, 288–289 (2009).

67. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).

68. Prossliner, T., Gerdes, K., Sørensen, M. A. & Winther, K. S. Hibernation factors directly block ribonucleases from entering the ribosome in response to starvation. *Nucleic Acids Res.* **49**, 2226–2239 (2021).

69. Battesti, A., Majdalani, N. & Gottesman, S. The RpoS-Mediated General Stress Response in Escherichia coli. *Annu. Rev. Microbiol.* **65**, 189–213 (2011).

70. von Wulffen, J., Sawodny, O. & Feuer, R. Transition of an Anaerobic Escherichia coli Culture to Aerobiosis: Balancing mRNA and Protein Levels in a Demand-Directed Dynamic Flux Balance Analysis. *PLoS ONE* **11**, e0158711 (2016).

71. van Elsas, J. D., Semenov, A. V., Costa, R. & Trevors, J. T. Survival of Escherichia coli in the environment: fundamental and public health aspects. *ISME J.* **5**, 173–183 (2011).

72. Antczak, M., Michaelis, M. & Wass, M. N. Environmental conditions shape the nature of a minimal bacterial genome. *Nat. Commun.* **10**, 3100 (2019).

73. Görke, B. & Stülke, J. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nat. Rev. Microbiol.* **6**, 613–624 (2008).

74. Weikert, C., Canonaco, F., Sauer, U. & Bailey, J. E. Co-overexpression of RspAB Improves Recombinant Protein Production in *Escherichia coli*. *Metab. Eng.* **2**, 293–299 (2000).

75. Chi, H. *et al.* Engineering and modification of microbial chassis for systems and synthetic biology. *Synth. Syst. Biotechnol.* **4**, 25–33 (2019).

76. Dobzhansky, T. Genetics of natural populations; recombination and variability in populations of Drosophila pseudoobscura. *Genetics* **31**, 269–290 (1946).

77. Breuer, M. *et al.* Essential metabolism for a minimal cell. *eLife* **8**, e36842 (2019).

78. Rocha, E. P. C. The replication-related organization of bacterial genomes. *Microbiology* **150**, 1609–1627 (2004).

79. Gaal, T. *et al.* Colocalization of distant chromosomal loci in space in E. coli: a bacterial nucleolus. *Genes Dev.* **30**, 2272–2285 (2016).

80. Lee, J. H. *et al.* Metabolic engineering of a reduced-genome strain of Escherichia coli for L-threonine production. *Microb. Cell Factories* **8**, 2 (2009).

81. Ma, S. *et al.* Random genome reduction coupled with polyhydroxybutyrate biosynthesis to facilitate its accumulation in Escherichia coli. *Front. Bioeng. Biotechnol.* **10**, 978211 (2022).

82. Aguilar Suárez, R., Stülke, J. & van Dijl, J. M. Less Is More: Toward a Genome-Reduced Bacillus Cell Factory for “Difficult Proteins”. *ACS Synth. Biol.* **8**, 99–108 (2019).

83. Shen, Y. *et al.* SCRaMbLE generates designed combinatorial stochastic diversity in synthetic chromosomes. *Genome Res.* **26**, 36–49 (2016).

84. Choi, J. *et al.* Precise genomic deletions using paired prime editing. *Nat. Biotechnol.* **40**, 218–226 (2022).

85. Jiang, T., Zhang, X.-O., Weng, Z. & Xue, W. Deletion and replacement of long genomic sequences using prime editing. *Nat. Biotechnol.* **40**, 227–234 (2022).

86. Ding, S. *et al.* Efficient Transposition of the piggyBac (PB) Transposon in Mammalian Cells and Mice. *Cell* **122**, 473–483 (2005).

87. Ivics, Z., Hackett, P. B., Plasterk, R. H. & Izsvák, Z. Molecular Reconstruction of Sleeping Beauty, a Tc1-like Transposon from Fish, and Its Transposition in Human Cells. *Cell* **91**, 501–510 (1997).

88. Voigt, F. *et al.* Sleeping Beauty transposase structure allows rational design of hyperactive variants for genetic engineering. *Nat. Commun.* **7**, 11126 (2016).

89. Tian, J. *et al.* Mage transposon: a novel gene delivery system for mammalian cells. *Nucleic Acids Res.* **52**, 2724–2739 (2024).

90. Morisaka, H. *et al.* CRISPR-Cas3 induces broad and unidirectional genome editing in human cells. *Nat. Commun.* **10**, 5302 (2019).

91. Wang, M., Sanfiorenzo, C., Zhang, R. J. & Wang, K. A universal system for streamlined genome integrations with CRISPR-associated transposases. 2022.05.30.494051 Preprint at https://doi.org/10.1101/2022.05.30.494051 (2022).

92. Meyer, A. J., Segall-Shapiro, T. H., Glassey, E., Zhang, J. & Voigt, C. A. Escherichia coli 'Marionette' strains with 12 highly optimized small-molecule sensors. *Nat. Chem. Biol.* **15**, 196–204 (2019).

93. Robertson, W. E. *et al.* Creating custom synthetic genomes in Escherichia coli with REXER and GENESIS. *Nat. Protoc.* **16**, 2345–2380 (2021).

94. Al'Abri, I. S., Haller, D. J., Li, Z. & Crook, N. Inducible directed evolution of complex phenotypes in bacteria. *Nucleic Acids Res.* **50**, e58 (2022).

95. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).

96. BBMap. *SourceForge* https://sourceforge.net/projects/bbmap/ (2025).

97. Shimoyama, Y. pyCirclize: Circular visualization in Python [Computer software]. (2022).

98. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).

99. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187-191 (2014).

100. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).