Understanding the metabolic burden of protein overproduction in *Escherichia coli* using multi-step directed evolution

Ingrid Afrodyta Herdzik
PhD

University of York
Biology

April 2024

Abstract

The recombinant protein production market has been growing since its advent in the 1980s with the introduction of the first recombinant protein-based pharmaceutical, insulin. Today, recombinant proteins are crucial not only in the biopharmaceutical industry but are widely used in everyday products such as cosmetics and detergents. Out of many available hosts for recombinant protein production, including yeast, plant, insect, and mammalian cells, bacterial hosts remain favoured by many manufacturers. *Escherichia coli*, one of the most common bacterial host choices for recombinant protein production, is a well-studied organism with ample techniques available for further genetic modification. However, producing heterologous proteins in *E. coli* comes with unique challenges; its rapid evolution and adaptation to stressful environments makes it an unsuitable candidate for some biomanufacturing processes, such as continuous fermentation. The work presented here not only showcases some of the evolutionary metabolic burden escape mechanisms employed by plasmid-carrying *E. coli* under high metabolic stress of recombinant protein production, but also introduces a novel Fluorescent Assisted Cell Sorting (FACS) screening protocol that can direct the experimental evolution of plasmid-carrying *E. coli* cells towards a more stable and more productive phenotype. Next-generation long-read sequencing (Oxford Nanopore) was used to identify some of the genetic changes associated with these novel evolved phenotypes. One such gene, pcnB, encoding a PolyA polymerase was found to accumulate discrete mutations altering the N- or C-terminal halves of the protein that drive potentially opposing phenotypes for productivity of recombinant protein. These results not only indicate that the modulation of PcnB function affects the productivity of recombinant protein-producing strains, but also highlight this gene as a potential target for modification in industrial *E. coli* strains, thus providing valuable insights for the biopharmaceutical and biomanufacturing industries.

Acknowledgements

This research project could not have been completed without the extensive network of support I could rely on throughout my journey. I would like to express my deepest gratitude to the following people:

My supervisors at University of York, Professor Gavin Thomas and Professor Ville Friman, who both cultivated my scientific curiosity and guided me while showing kindness and an understanding of my circumstances.

My Thesis Advisory Panel members, Professor Marjan van der Woude and Professor Daniela Barillà, who both played a major part in my development as a scientist and sparked my love for microbiology, as well as ensured that this research project progressed in a timely and meaningful fashion.

My industrial supervisors, Doctor Christopher Lennon and Sarah Ryan, who provided me with a deeper understanding of the inner workings of the biotechnology industry.

I would also like to extend my sincere thanks to the following:
The members of the Thomas lab group at the University of York, for providing me with an amazing environment in which I could grow both as a person and as a scientist.

I am also thankful to my family:
My fiancé, whose love and support were essential to my wellbeing and continued investment in my research.
My father, who cultivated my curiosity since I was a little girl and supported me in my pursuit of knowledge far from my home country.
My mother, who celebrated my journey and pushed me towards scientific excellence every day.

Author’s declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

Contents

List of tables

List of figures

1. Introduction

1.1 Recombinant protein production industry

The global recombinant protein production market is growing, with its value (revenue) estimated to reach 2.4 billion USD by 2027 (MarketsandMarkets 2022). From use in the cosmetic industry (de Oliveira *et al.* 2016; Brämer *et al.* 2019; Basit *et al.* 2018), food processing (Alvarez-Sieiro *et al.* 2014; Sheng *et al.* 2015) and detergents (Rajput *et al.* 2011; Gulmez *et al.* 2018), to pharmaceutical components (Govender *et al.* 2020; Martínez *et al.* 2012), the potential applications of recombinant proteins are vast. Therefore, many methods have been developed to improve the efficiency and reduce the associated costs of producing recombinant proteins.

In 1973 the first synthetic plasmid was constructed from fragments of existing plasmids digested with restriction enzymes which were then ligated and transformed into *Escherichia coli* (Cohen *et al.* 1973). This discovery was followed by the introduction of eukaryotic gene coding sequences into the plasmids (Morrow *et al.* 1974), however these lacked evidence of functional protein expression. The first functional eukaryotic proteins expressed from recombinant plasmids in *E. coli* were yeast and mold proteins (Vapnek *et al.* 1977; Struhl *et al.* 1976; Ratzkin & Carbon 1977). The first description of biologically active mammalian protein was mouse dihydrofolate reductase expressed in recombinant *E. coli* in 1978 (Chang *et al.* 1978). In 1979 the first human recombinant protein (insulin) was produced in *E. coli* (Goeddel *et al.* 1979) and three years later recombinant insulin was marketed as the first pharmaceutical containing recombinant DNA (Quianzon & Cheikh 2012) marking the start of the era of molecular biotechnology.

While *E. coli* and its plasmids were the first recombinant protein production platform developed, it was quickly followed by developments in protein production in

yeast. *Saccharomyces cerevisiae* was the first yeast host used to express human interferon gene in 1981 (Hitzeman *et al.* 1981). Heterologous protein expression in yeast quickly gained popularity (Buckholz & Gleeson 1991) due to yeast combining some of the main advantages of the bacterial expression systems (rapid growth, ease of culture, genetic manipulation tools available) with additional benefits (such as glycosylation of the products). With advancements in understanding of the glycosylation processes it is now well established that yeast protein glycosylation patterns are different from those of human proteins (Brooks 2004), and this has allergenic potential (Çelik & Çalık 2012).

Even though many other host alternatives are available for the recombinant protein production, *E. coli* remains one of the most popular choices. All hosts have their unique challenges and there are disadvantages to using them in certain applications; these are briefly summarised in **Table 1.1.** Since *E. coli* was the first host used for heterologous protein production, it is the most well studied, and there are many tools readily available to facilitate genetic modifications. Furthermore, its cultivation is cost efficient, and protocols are straightforward. Many of the originally identified challenges to protein production in *E. coli* have since been addressed, further increasing its attractiveness as a recombinant protein production host. For example, the glycosylation of the products in *E. coli* was first made possible by transferring the glycosylation system from *Campylobacter jejuni* to *E. coli* (Wacker *et al.* 2002) and since then approaches to improved glycan biosynthesis in *E. coli* have been described (Ding *et al.* 2017; Glasscock *et al.* 2018; Feldman *et al.* 2005); although this is still not widely applicable. Difficulties with correct folding of some heterologous proteins expressed in *E. coli* have been addressed by developing a new strain capable of disulfide bond formation (Lobstein *et al.* 2012). Secretion of the

recombinant proteins can be enhanced by using mutant strains with increased cell membrane permeability (Fordjour *et al.* 2023), employing specific cultivation protocols (Yim *et al.* 2001), and incorporation of secretion signals in the coding sequence (Choi *et al.* 2000; Low *et al.* 2013).

Today, the most popular therapeutic recombinant proteins are monoclonal antibodies (mAbs), produced mainly in mammalian cell cultures (Li *et al.* 2010). Monoclonal antibodies have applications in treatment of many difficult to treat diseases. For example, their specificity provides an advantage over traditional cancer treatments such as chemotherapy or radiotherapy (Zahavi & Weiner 2020). They also have implications in treatment of autoimmune conditions such as rheumatoid arthritis (Behrens *et al.* 2015) and viral conditions such as Chikungunya (Fric *et al.* 2012). Even though mAbs are predominantly produced in mammalian cell cultures, recent developments showcase the possibility of antibody fragment and full length antibody production in *E. coli* (Gaciarz *et al.* 2016; Lénon *et al.* 2020; Rashid 20*22)*.

Host	Main advantages	Main disadvantages
Bacteria	Variety of genetic manipulation tools available Ease of culture and cost-effectiveness High yields	Often the product requires extensive downstream processing (purification, addition of post-translational modifications)
Yeast	High yields Rapid and cost-effective Post-translational modifications	Glycosylation patterns vary and can be different from the target pattern
Filamentous fungi	Long-term genetic stability Ability to secrete produced proteins	Low yields for some secreted proteins
Insect cells	High yields Post-translational modifications Flexible target protein size	Incorrect or variable glycosylation patterns of expressed proteins
Mammalian cells	Accurate post-translational modifications Appropriate folding	Expensive Poor secretion
Transgenic plants	High yields Appropriate folding Accurate post-translational modifications Cost effective	Risk of contaminants such as pesticides
Transgenic animals	High yields Appropriate folding Accurate post-translational modifications	Expensive Ethical considerations Slow System complexity

Table 1.1. The comparison of main advantages and disadvantages of available recombinant protein production hosts. The table was constructed based on reviews (Yin *et al.* 2007; Demain & Vaishnav 2009).

1.2 Recombinant protein production vector components

Plasmid choice is as important as the host choice in designing a successful recombinant protein production platform. All plasmids follow the same general design, which includes several crucial elements **(Fig. 1.1)**. The origin of replication (*ori*), responsible for plasmid copy number; regulatory components such as operator, promoter, and transcription terminator; selection markers responsible for plasmid maintenance in the population; and finally, the coding sequence for the gene of interest (GOI). Some examples of these plasmid design elements relevant to *E. coli* will be discussed below.

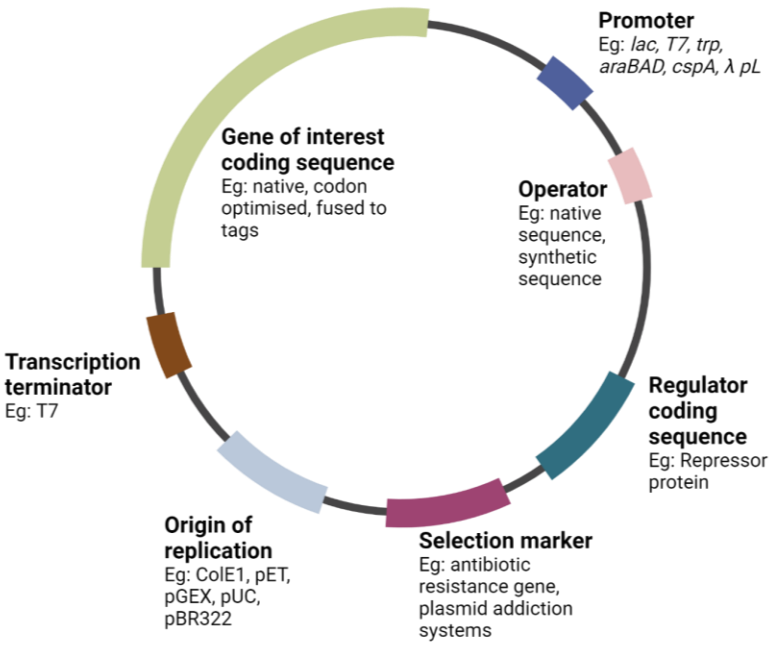


Fig. 1.1 The generic design features of a recombinant protein expression vector. The lack of directional information for each plasmid element is intentional, as the direction of the elements relative to each other is a part of a vector design process. Figure created with biorender.com

1.2.1 The origin of replication

The production of recombinant protein begins with the DNA sequence encoding the desired protein. The number of gene copies per host cell (gene dosage) is the first consideration before attempting to optimise the processes of transcription, translation, correct folding and post-translational modifications (PTMs). In *E. coli* transformed with a recombinant plasmid, the gene dosage is proportional to the plasmid copy number, although the relationship between gene dosage and protein product is not always linear (Uhlin & Nordström 1977).

The plasmid origin of replication *(ori)* is the determinant factor for the replication mechanism and plasmid copy number (PCN). Three main plasmid replication mechanisms exist - theta plasmid replication, strand displacement and rolling circle (del Solar *et al.* 1998). Most plasmids constructed for recombinant gene expression in *E. coli* are ColE1-like plasmids using theta plasmid replication mechanism (Tolmasky & Alonso 2015). The first three plasmids isolated and used for recombinant protein expression were pColE1, pMB1 and pSC101.

The ColE1 is a small (6,646 bp) plasmid, first isolated from *E. coli* and later shown to be responsible for colicin production (Bazaral & Helinski 1968; Tyler & Sherratt 1975). ColE1 was the first plasmid to be used in recombinant protein expression experiments (Hershfield *et al.* 1974), followed by independently isolated pMB1 plasmid which shared regions of homology and mechanism of replication with ColE1 (Bhagwat & Person 1981).

In ColE1-like plasmids, replication and PCN are determined by non-coding RNA sequences. The control of replication starts 555 bp upstream of the origin of replication, with the transcription of an RNA pre-primer (RNA II). RNA II forms a hybrid with DNA at the *ori* plasmid sequence and is subsequently extended by Polymerase I,

initiating the plasmid replication (Itoh & Tomizawa 1980). The plasmid copy number is modulated by the transcription of a 108 bp antisense RNA I. This RNA I forms a hybrid with pre-primer RNA II, making RNA II - DNA hybridization impossible (Cesareni *et al.* 1991). The original copy number maintained in a ColE1 plasmid is 25-30 (Hershfield *et al.* 1974), a medium copy number. This is a shared feature with pMB1 and its derivative pBR322 (Lin-Chao & Bremer 1986). A single nucleotide change in pMB1 *ori,* which gave rise to the pUC series of vectors, changed the PCN to several hundred by altering the secondary structure of RNA II impeding its interaction with RNA I (Lin-Chao *et al.* 1992; Anindyajati *et al.* 2016).

One of the beneficial features of medium and high copy number plasmids such as pBR322 and pColE1 is that they do not need to encode an active partitioning system. Instead, multiple copies of the plasmid are randomly distributed across the cell and inherited by the daughter cells. In addition, ColE-1-derived plasmids often contain a *cer* sequence, which is a resolvase target and aids in resolving plasmid multimers (Summers & Sherratt 1984), thus contributing to an even distribution of the plasmids to the daughter cells. A high PCN is also often desired to maximise the recombinant protein yield.

Plasmids with low PCN are also sometimes used in recombinant gene expression, as in contrast to high PCN plasmids, they pose a smaller metabolic burden to the host cell and, under some conditions, can be just as productive as high PCN plasmids (Jones *et al.* 2000; Peretti *et al.* 1989; Carrier *et al.* 1998). The first low copy number plasmid to be isolated and used in recombinant protein expression experiments was pSC101, isolated by controlled shearing of larger plasmid DNA (Cohen & Chang 1973). Plasmids with pSC101 *ori* have a low plasmid copy number and contain the *par* region, responsible for plasmid partitioning and stable inheritance

of the plasmids by the daughter cells (Manen *et al.* 1991). The need for an active partitioning system to maintain the plasmid in the population is one of the drawbacks of small PCN plasmids, as it uses the cellular resources in order to ensure plasmid inheritance and maintaining its copy number. This is an additional burden for the plasmid host

While traditional plasmid origin of replication sequences result in a specific PCN with small variation, therefore requiring consideration for specific applications, more recently, plasmids with tunable PCN have been constructed (Joshi *et al.* 2022; Rouches *et al.* 2022).

1.2.2 The selection marker

The next development in recombinant protein expression science, which started with the plasmids pColE1, pMB1 and pSC101, was to construct derivatives of those plasmids with improved characteristics. One of the desirable traits was the ability to efficiently screen for plasmid-carrying clones. To this end, a plasmid carrying two antibiotic resistance genes was constructed (pBR322). To enable positive selection its design included tetracycline resistance gene and ampicillin resistance gene with a *PstI* restriction enzyme site allowing for insertion of recombinant sequences into the plasmid (Bolivar *et al.* 1977; Bolivar*et al.* 1977).

The use of an antibiotic resistance gene on the plasmid and the addition of antibiotic to the growth media has become the most commonly used positive-selection method for plasmid retention. Producing heterologous proteins is metabolically and energetically costly for the bacteria and can lead to problems in plasmid maintenance in the population (San Millan & MacLean 2017; San Millan *et al.* 2018; Carroll & Wong 2018). This fitness cost can manifest itself as a reduced growth rate and a

disadvantage for bacteria carrying the plasmid when competing with plasmid-free isogenic bacteria; in the absence of positive selection, this may result in a complete plasmid loss from the population (San Millan & MacLean 2017; Subbiah *et al.* 2011). However, when antibiotics are used at scale it creates wastes that result in the release of antibiotics into the environment (Tang *et al.* 2021; Wang *et al.* 2020). Therefore, reducing or eliminating antibiotics used in industrial protein production is desirable.

One of the alternatives to antibiotic use are plasmid addiction systems. They work by ensuring that plasmid loss is toxic to the host cell (Kroll *et al.* 2010; Zielenkiewicz & Cegłowski 2001). Plasmid addiction systems can be divided into three main types: toxin-antitoxin system, metabolism-based addiction and operator repressor titration system (Kroll *et al.* 2010).

Toxin-antitoxin systems include coding sequences on the plasmid for a stable toxin and unstable antitoxin **(Fig. 1.2A)**. During cell growth, the plasmid-carrying cells produce both toxin and antitoxin, which interact with each other and prevent toxin function (Jurėnas *et al.* 2022). Unlike the stable toxin, which can be inherited by the plasmid-free daughter cells, the unstable antitoxin has to be produced from the plasmid directly in each cell to have an effect. In plasmid-free cells, the toxin interacts with its target (this varies from DNA replication through protein translation and cell integrity (Kroll *et al.* 2010; Jurėnas *et al.* 2022)). This interaction is lethal and results in cell lysis.

Both operator repressor titration and metabolism-based addiction systems require additional genetic manipulation of the host. In metabolism-based systems, an essential metabolic pathway gene is knocked out, which is then complemented by a copy included on the plasmid **(Fig. 1.2B)** (Kroll *et al.* 2009; Kroll 2010)). Meanwhile, in operator repressor titration systems, an essential gene is placed under the control of

an operator, which binds the repressor in plasmid-free cells, leading to cell lysis **(Fig. 1.2C)**. In plasmid-carrying cells, additional operator sequences are included on the multicopy plasmid. These additional operators bind the repressor molecules, sequestering them away from the operator upstream of the essential gene, leading to its derepression and cell survival (Cranenburgh *et al.* 2004).

Antibiotic resistance remains one of the most popular selection markers in recombinant protein production as, unlike some plasmid addiction systems, it does not require a specific, genetically modified host. This ensures vector versatility. However, other avenues for ensuring plasmid stability are being explored, such as improving plasmid segregation during cell division, discussed in this introduction's plasmid stabilisation approach section.

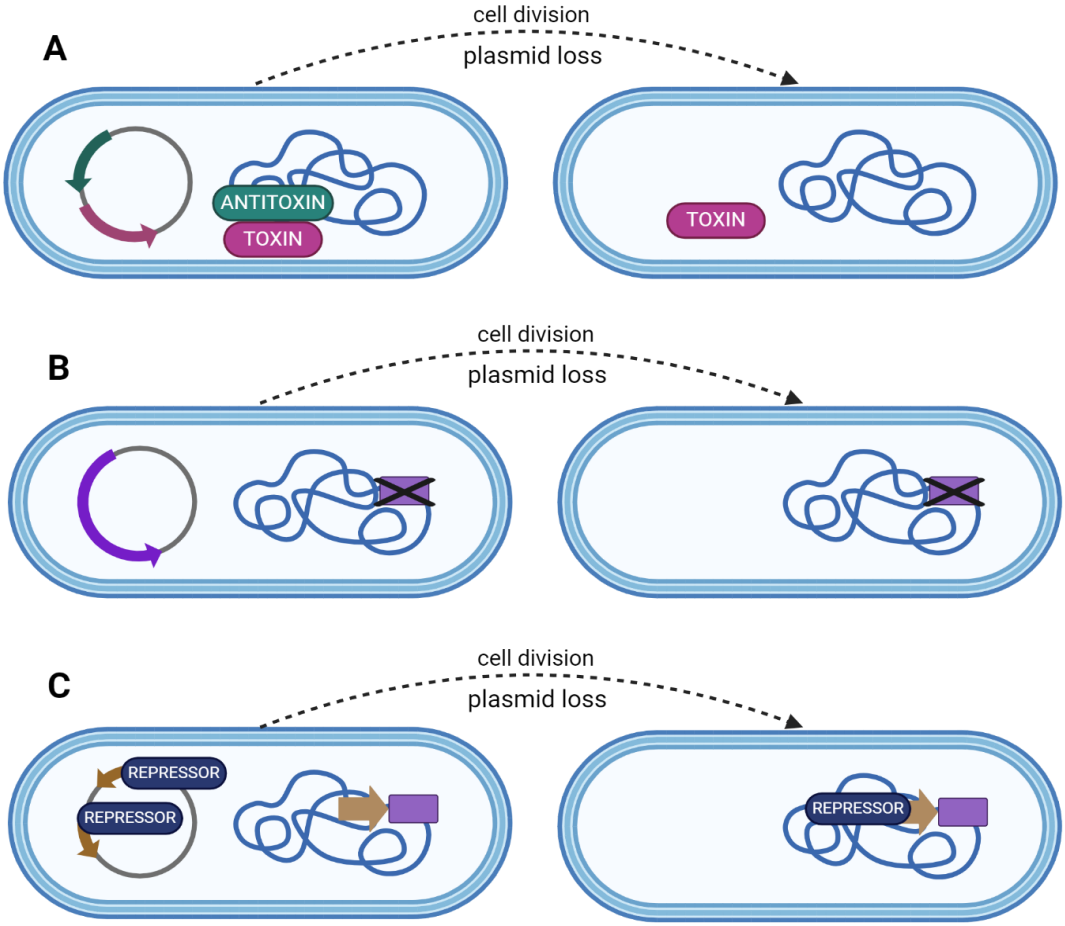


Fig. 1.2 Schematic representation of three plasmid addiction systems. A) Toxin-antitoxin system. A stable toxin and an unstable antitoxin are encoded on a plasmid. They interact with each other preventing the toxin from interacting with its target. In the event of plasmid loss, the stable toxin persists leading to cell death. **B)** Metabolism-based system. An essential gene is knocked out and complemented with a copy encoded on a plasmid. In the event of plasmid loss, the gene knockout is fatal. **C)** Operator repressor titration system. An essential gene is placed under control of an operator. The same operator sequences are present on the plasmid and sequester the repressor molecules away from the operator upstream of the essential gene. In the event of plasmid loss, the repressor binds to that operator and represses the expression of the essential gene, which is fatal to the cell. Figure created with biorender.com

1.2.3 Transcription regulation elements

Constitutive expression of the heterologous genes is not desirable, mainly because it imposes a significant metabolic burden on the host cell, which can lead to the selection of plasmid-free cells and a decrease in overall yield (San Millan & MacLean 2017; Neubauer *et al.* 2003). Therefore, in most plasmid-based production platforms, the gene of interest is expressed from an inducible promoter (controlled by an operator). The next generation of heterologous protein expression vectors were created to optimise the foreign gene expression from the plasmid.

The pET expression system was the first recombinant protein expression system designed to direct the heterologous protein expression from a plasmid in *E. coli.* This system is based on the T7 phage polymerase, which is highly selective for the T7 promoter. Therefore, when genes are placed on a plasmid under the control of a T7 promoter, they can only be transcribed upon introduction of the T7 polymerase into the host cell. While infecting the cells carrying a plasmid with a T7 promoter with T7 phage led to the expression of plasmid DNA controlled by the promoter, the RNA did not accumulate at sufficient levels before the infection killed the host (McAllister *et al.* 1981). Hence, the pET system was designed to introduce the T7 polymerase into the host using the lambda cloning vector D69 (Studier & Moffatt 1986). Among several tested options, the lysogen DE3 was tested, which introduced the T7 polymerase gene under the control of *lacUV5* promoter into the host genome. This allowed for the induction of T7 polymerase expression by IPTG (Studier & Moffatt 1986), however, it was also found that lacUV5 is only partially repressed in the absence of an inducer (Studier & Moffatt 1986). More recently, attempts have been made to enhance the control of phage polymerase expression by placing it under the control of other

inducible promoters (Du *et al.* 2021). The main disadvantage of pET expression system is that it requires the T7 polymerase encoded separately from the expression vector. While usually this requires a specific bacterial host, most commonly *E. coli* BL21(DE3), studies have shown that it may also be encoded on a second plasmid (Tabor 1990).

Sugar-inducible promoters such as *lac* and *araBAD* are other examples of inducible promoters requiring a specific inducer addition into the growth media to initiate recombinant protein production. They are well-characterised and widely used (Marschall *et al.* 2016). In these systems, plasmids encode repressor proteins, which interact with the matching operator DNA sequences, blocking RNA polymerase interaction and plasmid gene transcription. In the presence of the specific sugar, the repressor protein dissociates from the operator, allowing gene expression. For lactose inducible system, there have been gratuitous inducers identified (Isopropyl ß-D-1-thiogalactopyranoside, IPTG and methyl-1-thio-β-d-galactopyranoside, TMG), which allows for a single dose of inducer in a fermentation process. The host cannot metabolise gratuitous inducers, and once added they continue to de-repress heterologous protein production until the end of the growth process (Marbach & Bettenbrock 2012).

Sugar availability is not the only stimulus useful in the design of inducible recombinant protein expression systems. Some inducible promoters respond to environmental changes such as temperature (both low (Bartolo-Aguilar *et al.* 2022) and high (Han *et al.* 2004)) or pH (Chou *et al.* 1995). Furthermore, any inducible promoter system described above can be further modified and refined to improve control over the heterologous protein expression. Elements from several systems can be combined to create entirely new systems, such as *tac* promoter, a hybrid between

lac and *trp* promoters (de Boer *et al.* 1983), which shows enhanced productivity compared to the original promoters. Modifying the operator sequence can enhance repressor binding and reduce heterologous protein expression in the absence of an inducer. An example is the perfect palindromic lac operator (Sadler *et al.* 1983; Simons *et al.* 1984).

1.2.4 Gene of interest sequence

The final component of a successful recombinant protein production vector design is the coding sequence of the gene of interest (GOI). Producing a specific protein requires a strictly predefined sequence of amino acids. However, genetic code is redundant, and therefore, one amino acid can be encoded by multiple codons. Moreover, many species exhibit codon bias, a preference for encoding a specific amino acid with a specific codon in their native proteins. This often correlates with the abundance of corresponding tRNAs (Parvathy *et al.* 2021). Codon optimisation is one of the strategies used to address a disparity between heterologous protein codon usage and host codon bias, where the sequence of the GOI is altered (Al-Hawash *et al.* 2017; Boël *et al.* 2016). Codon optimisation issues can be mitigated through supplementation of rare tRNAs on a separate compatible plasmid (such as in case of Rosetta(DE3) strain (Burgess-Brown *et al.* 2008; Zhong *et al.* 2017)).

In addition to the core protein coding sequence, in some cases proteins may be expressed as fusions, which aids in downstream processing and product detection. One of the vector families designed to allow the expression of fusion proteins are the pGEX vectors (Smith & Johnson 1988; Hakes & Dixon 1992; Berthold *et al.* 1992; Guan & Dixon 1991). Specific fusion tags may also be used to enhance solubility of the target protein (Esposito & Chatterjee 2006).

1.3 Recombinant protein production host choice

The host choice is a crucial step in designing a successful heterologous protein production protocol. For microbial fermentation, *E. coli* remains the popular choice due to its ease of cultivation, cost-effectiveness, and well-characterised genetics. However, there are many *E. coli* strains, each with unique genetics impacting their performance in batch and fed-batch cultures. The two genetic backgrounds most relevant to this study are BL21 and K-12 derivative strains, also called B and K genetic backgrounds **(Fig. 1.3).** This is because those two strains are the main host choices for FujiFilm Diosynth Biotechnologies, the industrial collaborator of this project. FujiFilm is focused on delivering quality heterologous protein products to its customers in both biotechnology and biopharmaceutical sectors. This project presents an opportunity to optimise one of their existing protein production platforms – pAVEway.

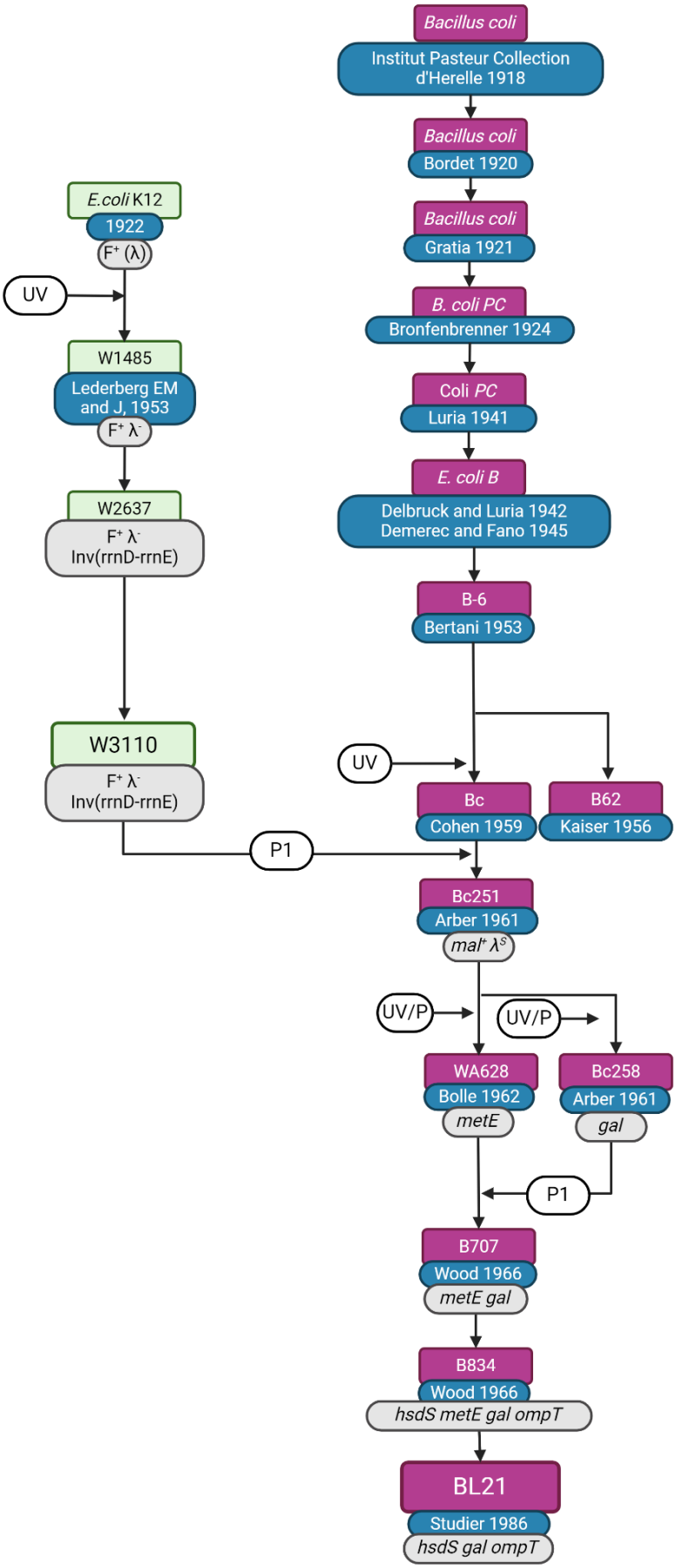


Fig. 1.3 The ancestry and relationships between *E. coli* K and B strains. The K genetic background strains are in green boxes, while the B strains are in magenta boxes. The blue boxes contain information about the date each strain was first isolated or described. The relevant genotypes are included in the grey boxes. P1 transductions, UV treatments and UV followed by penicillin selection are indicated in the white boxes. Figure based on (Daegelen *et al.* 2009) and created with biorender.

Comparing BL21 and W3110 (a K-12 strain) *E. coli* genomes available online (GenBank CP010816.1 and AP009048.1) reveals a difference in genome size. BL21 has a smaller genome (118 214 bp shorter) than W3110. The genomes are highly similar and closely related **(Fig. 1.3)**; however, each has some unique features. The BL21 strain genome lacks flagellar biosynthesis pathway genes (Studier *et al.* 2009), and its cell envelope composition is responsible for increased permeability compared to other strains (Herrera *et al.* 2002). These differences in cell envelope composition have been reported as causes for varied responses to stress (Yoon *et al.* 2012). The BL21 strain also does not encode the outer membrane protease OmpT, which is involved in degradation of protein fusions, negatively affecting the overall yield (Baneyx & Georgiou 1990). It also cannot consume galactose, making it a valuable host for some auto-induction protocols (Xu *et al.* 2012). Meanwhile, the lack of flagellar genes allows for resource conservation and is a desirable trait for a heterologous protein production host. In contrast, *the E. coli* K-12 genome contains genes responsible for the very short patch repair system (Yoon *et al.* 2012) (responsible for correcting GT mismatches), which can reduce the rates of mutation on the recombinant plasmids. This makes *E. coli* K strains valuable alternative plasmid hosts for the expression of recombinant proteins.

Traditionally, K background strains have been used in laboratory experiments such as cloning. This originated with the work of Douglas Hanahan and the studies on transformation efficiency in various *E. coli* strains (Hanahan 1983; Grant *et al.* 1990) which introduced *E. coli* DH5α. Meanwhile, following the pET expression system development, B background *E. coli* (BL21(DE3) specifically) became the predominantly used strain in heterologous protein production (Studier & Moffatt 1986). More recently, *E. coli* BL21 mutant has been adapted to produce stable plasmid DNA

molecules by inactivation of the same genes which are inactive in traditionally used K background strains - *endA* and *recA (*Phue *et al.* 2008*)*. Meanwhile, a K background strain HMS174(DE3), originally described in the same paper from which BL21(DE3) originates, has been found to be a suitable alternative host for recombinant protein production (Hausjell *et al.* 2018). These recent discoveries highlight the delicate interplay between plasmid and host strain interactions leading to specific phenotypes and explain why both K and B background strains are used in the protein production at scale at Fujifilm Diosynth Biotechnologies.

Examining the genome differences between the two *E. coli* genetic backgrounds provides a limited understanding of the driving factors causing their distinct phenotypes. While both strains may contain the same genes, those may be expressed at different levels. Transcriptomic analysis of *E. coli* B and K strains grown in high-density cultures reveals differences in the expression of essential metabolic genes. For example, differences in acetate accumulation in B and K *E. coli* cultures have been reported despite no differences in genes involved in the TCA cycle and glyoxylate shunt pathways between the strains. Transcriptome analysis revealed that several glyoxylate shunt pathway genes were expressed at higher levels in B background *E. coli* than in K in high density, high glucose cultures (isocitrate lyase and malate synthase), alongside Acetyl-CoA synthetase responsible for acetate metabolism (Phue & Shiloach 2004). Meanwhile, the isocitrate lyase repressor responsible for repressing the glyoxylate shunt pathway was expressed at higher levels in the K background strain than in B in the same experiment, alongside acetate kinase and pyruvate oxidase responsible for acetate accumulation (Phue & Shiloach 2004). These differences are significant during batch and fed-batch cultivations during

the heterologous protein production process, as high acetate concentration impairs growth and impacts biomass accumulation and overall yield (Phue *et al.* 2005).

Acetate management pathways are not the only metabolic pathways that differ in their gene expression levels between *E. coli* B and K. Microarrays and northern blot investigations have also revealed differences in glucose metabolism. Increased gluconeogenesis and glycogen synthesis gene transcription in *E. coli* B compared to K might be helping in acetate concentration reduction in high glucose concentration culture (Phue *et al.* 2005).

One limitation of transcriptome studies is that they assume that transcription level directly correlates with protein content and activity of the gene products. This is not always accurate, as translation can be affected by codon and tRNA availability (Olivares-Hernández *et al.* 2011) alongside varying mRNA stability and interactions with non-coding RNAs (Nilsson *et al.* 1984; Belasco *et al.* 1986; Storz *et al.* 2004). In order to understand the transcriptome data, it is essential to interpret it in the context of proteome data. In *E. coli* grown in high-density cell cultures, transcriptome and proteome profiles have been found to follow very similar trends (Yoon *et al.* 2003). It is also important to note that transcriptome and proteome are not a static quality of cultures; they are changeable and can be influenced by environmental factors such as stress caused by induction (Dürrschmid *et al.* 2008). Induction of recombinant protein expression results in upregulation of chaperone and aggregation sensor proteins such as ClpB and IbpA. Moreover, the changes in transcriptome and proteome of the host cell expressing recombinant protein heavily depend on the metabolic burden imposed by specific products (Dürrschmid *et al.* 2008).

E. coli B and K strains also differ in their extracellular proteome, the proteins released into the media during high cell density cultivation. *E. coli* B can release more

and bigger proteins than *E. coli* K, most likely due to differences in the outer membrane permeability caused by different levels of several porin proteins; meanwhile, *E. coli* K was found to release more cytoplasmic proteins (Xia *et al.* 2008). These differences can be exploited when choosing the host organism best suited for expressing specific target proteins.

All of the differences between *E. coli* B and K strains at the genome, transcriptome and proteome levels contribute to vastly different phenotypes and their utility in heterologous protein production. The choice of the host genetic background for protein production depends on the target protein to be produced and may require expression trials before being scaled up for production.

1.4 Fermentation and evolutionary pressure

1.4.1 Fermentation protocols

Once both the host and the plasmid have been chosen, the final step is to design a fermentation protocol. The main traits of a successful fermentation protocol are cost-effectiveness and final product yield.

Three main designs of fermentation protocols are used in recombinant protein production: batch, fed-batch and continuous fermentation. In batch fermentation, the starting media contains all the components the culture will use to grow. Fed-batch fermentation is an improved process where the total biomass is enhanced by adding feed solution to the fermenter at a specified rate until the final fermentation volume (Hewitt *et al.* 1999). Continuous fermentation further enhances the total biomass achievable and total protein yield by constant flow of fresh feed solution and simultaneous product removal (Li *et al.* 2014), making it an attractive alternative to batch and fed-batch protocols.

Fed-batch fermentation is most often used for production with *E. coli* as the host. Use in continuous production is not feasible due to plasmid-free cells gaining fitness advantage over plasmid-carrying cells, which in the absence of selection results in rapid plasmid loss at the population level and a sharp decrease in recombinant protein yield (Sieben *et al.* 2016; Vyas *et al.* 1994). The current understanding of the cause of the substantial fitness cost of maintenance of plasmids reveals several possible mechanisms, and a combination of any of those factors is likely present in any host and plasmid combination (San Millan & MacLean 2017). Of those, three are of interest concerning this study: the cost of replication of a multicopy plasmid, often related to plasmid replication sequestering the host replication

machinery away from the genome and slowing cell division (Carroll & Wong 2018); the cost of expressing high levels of plasmid-encoded genes (such as expression from a strong promoter T7A3 in pAVE011), related to amino-acid starvation (San Millan & MacLean 2017); and specific genetic conflicts arising between the host and the plasmid which are more difficult to predict (Hall *et al.* 2021).

1.4.2 Evolutionary pressure during fermentation

Bacteria containing plasmids have been shown to undergo evolutionary processes to escape the burden of heterologous protein production (Smith & Bidochka 1998; James *et al.* 2021; Million-Weaver *et al.* 2012; Modi & Adams 1991). Although the most common solution is to discard the plasmid entirely, sometimes single targeted mutations can be involved on either the plasmid, the bacterial chromosome, or both (Yano *et al.* 2016; Millan *et al.* 2015). These compensatory mutations may alleviate the cost of plasmid carriage over time in a co-evolution process and could potentially improve plasmid-based protein production in bacterial bioreactors. The most widely studied examples of the coevolutionary processes involve antibiotic resistance plasmids. Antibiotic resistance in clinical settings has been identified as a serious threat in recent years (Abushaheen *et al.* 2020; Huemer *et al.* 2020) and often the antibiotic resistant genes are carried on plasmids. Coevolution between plasmids and their hosts has been found to lead to an expanded host range and better fitness of the plasmid-carrying strains (De Gelder *et al.* 2008; Loftie-Eaton *et al.* 2015), posing a significant threat to human health. Once the plasmid is transferred into a new host, it usually imposes a significant carriage burden and the antibiotic resistance protein expression cost. Through coevolutionary processes, the plasmid and the host can then resolve any conflicts which stabilise the plasmid in the new host population (Bottery *et al.* 2017; Stalder *et al.* 2017). Altogether, the coevolutionary processes can facilitate

the emergence of novel multidrug resistant pathogens (Jordt *et al.* 2020; Benz & Hall 2023).

Despite the fact that plasmid-bacteria co-evolution is widely occurring, the impact of coevolutionary processes in industrial protein production strains remains understudied. The same features making *E. coli* an attractive recombinant protein production host - rapid growth, robustness in variety of growth conditions - are contributing to the mutational instability of the vectors. Given the approximates for: generation time of *E. coli* of 20 min (Son & Taylor 2021), batch fermentation of 12 h (Yang & Sha n.d.) and spontaneous mutation rate of 1 × 10^-3 per genome per generation (Lee *et al.* 2012), it is clear that in high volume, high density cultures random mutations can arise swiftly. Combined with high selective pressure against plasmid carriage (since producing heterologous proteins from a plasmid imposes a significant metabolic burden on the host) this could lead to the evolution of non-productive clones with a fitness advantage over the ancestor. Those non-productive clones could then overtake the population in the fermenter and cause a significant production efficiency loss.

The coevolutionary changes can affect various host and plasmid regions. For example, when the plasmid carriage cost results from the replication of multicopy plasmids, mutations may arise in the origin of replication (*ori*) region. These alter the plasmid copy number per cell, thus decreasing the plasmid gene expression and the number of plasmid copies synthesised in each cell division cycle (Million-Weaver *et al.* 2012). When the protein overexpression itself is a burden to the plasmid-carrying bacteria, point mutations targeting the specific polymerase (such as T7) reading the promoter region can arise (James *et al.* 2021). Distinct genetic conflicts between the plasmid and the host can be resolved with mutations in those conflicting DNA

sequences (Hall *et al.* 2021). Some of these mutations could improve the stability and protein expression, but this has not yet been explored.

1.4.3 Plasmid stabilisation for recombinant protein production

In order to prolong fermentation and improve the yields, plasmid stabilisation is crucial. There are two types of plasmid instability - segregational and mutational. The first can be addressed by ensuring the plasmids are inherited evenly across generations. One example is incorporating multimer resolution systems in plasmid design (Allen *et al.* 2022).

Mutational plasmid instability is more challenging to address. In *E. coli*, the mutation rate is relatively high (Wielgoss *et al.* 2011; Lee *et al.* 2012), and generation time is short (Wang *et al.* 2010; Son & Taylor 2021). This increases the chances of random mutations of the plasmid, which may increase the host strain's fitness. Combined with high plasmid carriage cost acting as an evolutionary pressure, the ancestral strain is outcompeted, and the mutated plasmid stabilises the population.

The solution to this problem might be reducing the likelihood of mutations providing evolutionary advantage to the host through initial coevolution. Coevolution is an interaction between the plasmid and the host, whereby random mutations accumulate over the generations on both the plasmid and host genome. These mutations then alleviate the cost of plasmid carriage and decrease the likelihood of further mutations providing fitness advantage and being selected for in a population (Bottery *et al.* 2018; Yano *et al.* 2016). However, in heterologous protein production the improved fitness often comes at a cost of reduced productivity.

1.4.4 Evolutionary microbiology

Experimental evolutionary microbiology is an l approach in which bacteria are grown in predefined conditions over multiple generations. One of the most well known experiments of this design is the Long Term Evolution Experiment (LTEE) with *E. coli,* which started in 1988 and was designed to assess differences between adaptations to the same environment in several cultures of the same strain (Lenski *et al.* 1991). This experiment led to better understanding of the dynamic evolutionary process and how genetic changes can become fixed in populations over thousands of generations (Maddamsetti *et al.* 2015). It was also possible to identify novel phenotypes within those evolved populations and explore the genetic changes responsible for them (Lamrabet *et al.* 2019).

LTEE did not involve distinct selective pressures designed to direct the bacterial evolution towards a specific phenotype; instead, it was investigating how population drift and random effects can affect the genetics of the evolved strains adapting to the environment. Including selection pressures in the design of evolutionary experiments allows for accumulation of mutations leading towards a desired phenotype. Through directed evolutionary experiments, it may be possible to select only for those mutations that improve the fitness of plasmid carrying strains without the productivity loss. The resulting host-plasmid pair may then be used as an improved protein production platform, less likely to be affected by coevolution processes during the fermentation itself. The evolutionary approach to improving the host-plasmid relationship for industrial protein production has the main advantage of not requiring any predefined targets which may aid in plasmid stabilisation. Traditionally modifications to existing expression platforms start with altered genotype and hypothesised phenotype based on the modified target. Instead, the evolutionary approach starts with the desirable

phenotype being selected for, and the associated genotype can be revealed through next generation sequencing.

1.5 Project aims

One of the platforms used in microbial fermentation is pAVEway by FujiFilm, using *Escherichia coli* as the bacterial plasmid host (Lennon 2014). There are several pAVEway plasmids used by FujiFilm, however this study will focus on pAVE011 and its derivatives. PAVEway plasmid pAVE011 design includes two palindromic *lac* operators on either side of a strong (T7A3) promoter. This unique feature allows for the precise control of the expression of the heterologous gene, as repressor binding induces DNA looping and prevents the target sequence from being transcribed by blocking the polymerase complex from accessing the promoter. PAVEway vectors and *E. coli* strains routinely used as their hosts (BL21 and W3110 *ΔompT*) will be used in this project to investigate the plasmid stabilisation potential of coevolutionary experiments. The goals of the study are:

1. To characterise the pAVEway vector and its impact on the host *E. coli*;
2. To guide the plasmid-host coevolution towards a more stable phenotype using a combination of engineering and evolutionary microbiology;
3. To understand the genetic changes responsible for the altered phenotype through next-generation sequencing of both the plasmid and the host.

In this thesis, each goal will be addressed in a separate results chapter. Firstly, the pAVEway-carrying *E. coli* strains will be assessed in terms of their growth and productivity using 96-well based assays in a plate reader. The growth will be estimated using OD600 measurements and the productivity using green fluorescence readings. Standard molecular cloning tools will be used to introduce any modifications to the plasmids before the evolutionary experiment. Secondly, during the evolutionary

experiment, fluorescent cell sorting will be used periodically to ensure sustained productivity in the evolved pAVEway carrying lines. Finally, next generation sequencing methods (Illumina and Nanopore) will be compared and used to determine the host and plasmid mutations responsible for the evolved phenotypes.

2. Methods

2.1 Bacterial strains, media preparation and incubation parameters

The list of strain genotypes and their names used in this study routinely is available in **Table 2.1**. FujiFilm Diosynth Biotechnologies (FDBK), Billingham, UK, provided two pAVE011 plasmid-harbouring strains, 1515 and 1516, along with non-transformed host cryo-stock of expression strains for this project. Laboratory stock of *Escherichia coli* DH5α was used primarily as the first host for transformation with new plasmid variants.

The media used throughout this study was LB-Miller (per 1L: NaCl 10g, tryptone 10g, yeast extract 5g), sometimes with vegetable tryptone (Millipore) instead of tryptone - it is then referred to as vegetable LB or vLB. Unless otherwise stated, antibiotics were added to select for plasmids at the following final concentrations: kanamycin 30 µg/ml (for the *recA::kan* strain), tetracycline 10 µg/ml (for pAVE011, pIAH011 and pCMT-FLP plasmids), chloramphenicol 30 µg/ml (for pAVE011CMR) and ampicillin 100 µg/ml (for pKD46 and pCP20). Unless otherwise stated, the final concentration of the inducer isopropyl ß-D-1-thiogalactopyranoside (IPTG) was 0.5 mM. Where induction was performed in the exponential growth phase, the IPTG was added after the cultures reached an OD_{600} >0.4 in a culture started from an overnight stock. This protocol is consistent with the methods used in-house by FDBK. Where induction was performed during the lag phase of growth, a cryopreserved vial sample was used to inoculate antibiotic and inducer-supplemented media directly.

All growth steps were at 37°C, unless growing a strain containing a temperature-sensitive plasmid such as pCP20, pKD46 or pCMT-FLP, in which case the incubation was at 30°C. All cultures were shaken at 180-220 rpm. Cultures were cured of the plasmid pKD46 by overnight incubation at 37°C and of the pCP20 and

pCMT-FLP plasmids by overnight incubation at 42°C following the protocol outlined in Datsenko & Wanner (Datsenko & Wanner 2000).

All cryostock preservation steps involved mixing a culture sample with 40% sterile glycerol in a 1:1 ratio. Cryostocks were stored in 1 ml aliquots in a freezer at -70 °C.

All temperature-sensitive media components such as antibiotics, IPTG and glucose solution were filter sterilised using 0.2 µm syringe filters and stored at -20°C or 2-8°C as appropriate.

Strain name used in this study	Genetic information	Notable details
1515	W3110 Δ*ompT* pAVE011 *sfgfp*	CLD1515 provided by FujiFilm. K background strain.
1516	BL21 pAVE011 *sfgfp*	CLD1516 provided by FujiFilm. B background strain
IAH015	W3110 Δ*ompT* pIAH011 *sfgfp*	K background strain. Transformed with a plasmid with modified operator 1
IAH016	BL21 pIAH011 *sfgfp*	B background strain. Transformed with a plasmid with modified operator 1.
830	W3110 Δ*ompT* pAVE011	CLD830 provided by FujiFilm. K background strain. Plasmid does not encode heterologous protein
1944	BL21 pAVE011	CLD1944 provided by FujiFilm. B background strain. Plasmid does not encode heterologous protein
RA015	W3110 Δ*ompT* Δ*recA* pAVE011 *sfgfp*	Strain 1515 with an additional recA deletion
RA016	BL21 Δ*recA* pAVE011 *sfgfp*	Strain 1516 with an additional recA deletion
DH5α	*recA1, endA1*	Strain routinely used for cloning. From University of York laboratory stock
Keio collection recA mutant	BW25113 *recA::kan*	Kanamycin resistance gene replaces *recA.* From University of York laboratory stock

Table 2.1. List of the *E. coli* strains used in this study. The table contains their names, genotype and other notable details.

2.2 Plasmid isolation and alteration

The square brackets in the below sequences indicate the O1' operator sequence. The following synthetic oligonucleotide containing relevant overhangs was used to insert the changed sequence:

5'-CCT AGA GGT CCC CTT TTT TAT TTT AAA ACC ATG TG**[**G AAT TGT TAT CCG GAT AAC AAT**]** TCA AGA ACA ATC CTG CAC GAA TTC AAA CAA AAC G-3'.

All plasmid isolations throughout the project were performed using an alkaline lysis plasmid miniprep kit (GenElute HP, Sigma Aldrich). The pAVE011 plasmid was altered using the Vazyme Clonexpress II kit using homology and recombination to replace the original operator 1 sequence with the alternative operator 1 sequence.

2.3 PCR and sequencing

List of PCR and sequencing primers used throughout the project can be found in **Table 2.2.**

Primer name	Primer sequence	Notable features
pram1F	5'-TCG TGGAAACGATAG GC-3'	Used with pram1R to amplify the promoter region in plasmids pAVE011 and pIAH011.
pram1R	5'-GTT TCA TGT GAT CCG GAT AAC G-3'	Used with pram1F to amplify the promoter region in plasmids pAVE011 and pIAH011.
pram2FG	5'- CCG GTC GTG CAG ATA AAC TCC-3'	Binds within the sfGFP gene. Used with pram2R to amplify the promoter

Primer name	Primer sequence	Notable features
		region in plasmids pAVE011 and pIAH011.
pram2R	5'- TGA TCG TGG AAA CGA TAG GC-3'	Binds upstream of O1. Used with pram2F to amplify the promoter region in plasmids pAVE011 and pIAH011.
pram2FC	5'-AGC TGA ACG GTC TGG TTA TAG G-3'	Binds within the chloramphenicol resistance gene. Used with pram2R to amplify the promoter region in plasmids pAVE011CMR and pIAH011CMR.
recAF	5'-CAG CAC TGG GCC AGA TTG AGA AAC-3'	Used with recAR to amplify a segment of *recA* gene which contains a point mutation in *recA1* strains
recAR	5'-GTA GTG GTT TCC GGG TTA CCG AAC-3'	Used with recAF to amplify a segment of *recA* gene

Primer name	Primer sequence	Notable features
		which contains a point mutation in *recA1* strains

Table 2.2. List of primers used in PCR and sequencing reactions throughout the project.

Primer pairs pram1 and pram2 **(Table 2.2)** are used in PCR reactions to amplify the promoter region of plasmids pAVE011 and pIAH011. Pram2FG, pram2FC and pram2R primers have been designed using Eurofins Sequencing Primer Design Tool (Eurofins Genomics n.d.) to enable their use in both PCR and sequencing reactions. When either pair is used in a PCR, the results are DNA fragments of distinct length (91 base pair difference) depending on whether a deletion occurred (shorter) or not (longer) in the promoter region. The anticipated product lengths for the intact and recombined promoter region are:

- 479 bp and 388 bp (pram1F and pram1R)
- 393 bp and 302 bp (pram2FG and pram2R)
- 360 bp and 269 bp (pram2FC and pram2R)

Colony PCR was used to amplify the *recA::kan* cassette from Keio collection strains. This was done by resuspending a colony from an agar plate in 10 µl of MiliQ water and using this as the DNA template in the PCR reaction setup. During cycling, the first denaturation step was extended to 10 minutes. Sequences of primers amplifying the *recA::kan* cassette are provided in the supplementary materials for the original paper (Baba et al. 2006).

In several experiments, a 600 bp fragment of the *recA* gene encompassing the single point mutation found in the *recA1* in DH5α, which deactivates the recombinase,

was amplified via PCR using a high fidelity polymerase (Phusion, NEB) and the primers recAF and recAR **(Table 2.2).**

All PCRs were performed using high fidelity polymerase (Phusion or Q5, NEB) whenever the products were to be sequenced and Go Taq G2 (NEB) in experiments not requiring sequencing. The protocols used were the default protocols, as advised on the polymerase manufacturer's website, with following thermocycling conditions:

- Promoter region amplification using Q5 polymerase:
    - Initial denaturation, 98°C for 30 seconds
    - 35 cycles: 98°C for 10 seconds, 65°C for 30 seconds and 72°C for 7 seconds
    - Final extension, 72°C for 2 minutes
- Promoter region amplification using Go Taq G2 polymerase:
    - Initial denaturation, 95°C for 2 minutes
    - 25 cycles: 95°C for 1 minute, 58°C for 15 seconds, 72°C for 30 seconds
    - Final extension, 72°C for 5 minutes
- Insert amplification from Keio collection strains using Phusion polymerase:
    - Initial denaturation, 98°C for 3 minutes
    - 35 cycles, 98°C for 10 seconds, 72°C for 36 seconds
    - Final extension,72°C for 10 minutes
- Entire *recA* region amplification using Go Taq G2 polymerase:
    - Initial denaturation, 95°C for 2 minutes
    - 35 cycles, 95°C for 1 minute, 59°C for 1 minute, 72°C for 1minute 50 seconds
    - Final extension, 72°C for 5 minutes
- Entire *recA* region amplification using Phusion polymerase:

    - Initial denaturation, 98°C for 30 seconds
    - 35 cycles, 98°C for 10 seconds, 65°C for 30 seconds, 72°C for 40 seconds
    - Final extension, 72°C for 2 minutes
- Mutation region of interest from *recA1* variant amplification using Q5 polymerase:
    - Initial denaturation, 98°C for 30 seconds
    - 35 cycles, 98°C for 7 seconds, 65°C for 30 seconds, 72°C for 20 seconds
    - Final extension, 72°C for 10 minutes.

In addition, all PCR products were purified using the NucleoSpin PCR clean-up kit. Isolated DNA purity assessment and quantity estimation were performed with Nanodrop, and whenever the experimental design demanded high accuracy, Qubit High-Sensitivity DNA assay was performed.

Short DNA fragments were sequenced in both forward and reverse directions, using LightRun Tube Eurofins Genomics sequencing service, and the results were analysed using benchling.com online tools.

2.4 DNA transformation methods

The *E. coli* cells were made either chemically competent or electrocompetent. For either method, the competent cells were obtained by inoculating 20 ml of LB media with 1 ml of an overnight culture. The culture was then incubated for approximately 1 hour at either 30 or 37℃, with shaking. Next, 1 ml of this culture was aliquoted into Eppendorf tubes. The following steps were all performed on ice, including incubations, and all solutions used were ice-cold and filter-sterilised.

For chemically competent protocol, the tubes were incubated for 1 hour and then centrifuged at 6000g, 4°C for 5 minutes. The resulting pellets were resuspended in 100 µl of 100 mM calcium chloride solution and incubated for 20 minutes on ice. Finally, the cells were centrifuged as above and resuspended in a final volume of 50 µl of 100 mM calcium chloride solution.

Electrocompetent cells were obtained by pelleting the cultures immediately after placing them on ice using centrifuge settings as above. These cells were washed four times with 10% glycerol. During each wash, the pellet was resuspended in progressively smaller volumes of glycerol (starting from 1ml) and pelleted again by centrifugation at 6000 rcf, 4℃ for 3 minutes.

The competent cells (prepared with either method) were mixed gently with the transformation DNA in the following proportions: for each 100 µl of competent cells, approximately 200 ng of the DNA was added.

The chemically competent cell and DNA mixture was then incubated on ice for 30 minutes. The cells were then heat shocked at 42°C and immediately placed on ice for 2 minutes. Next, 900 µl of the recovery media (LB or SOC) was added. The electrocompetent cell and DNA mixture was incubated on ice for 1 min. It was then transferred to an ice-cold 2 mm electroporation cuvette, and electroporation was performed using BioRad MicroPulser and its Ec2 programme. 1 ml of recovery media (vLB or SOC) was immediately added to the cuvette and mixed. The resulting culture was transferred into a 2 ml Eppendorf tube.

The transformed culture (from either protocol) was placed on a stable surface (not shaking) at either 30 or 37°C for 10 minutes to allow cells to recover. The tube was then moved to a shaker and left for 45 minutes to 1 hour. Following incubation, the cells were pelleted in a room-temperature centrifuge (6000 rcf for 3 minutes) and

resuspended in 400 µl of media. 100 µl of these resuspended cells were then plated onto selection LB agar plates, supplemented with an appropriate antibiotic. Sterile glass beads (5-10 per plate) were used to spread the bacteria evenly through vigorous shaking, directing the beads across the plates rather than around their edge.

The *recA* gene knockout in *E. coli* BL21 and W3110 *ΔompT* was performed using the λ Red recombination technique (Datsenko & Wanner 2000). Briefly, the strains were transformed with pKD46 plasmid, induced with arabinose (final concentration 2% w/v) and transformed with the *recA::kan* PCR product. They were then cured of the pKD46 plasmid and transformed with pCP20 or pCMT-FLP plasmid. These plasmids encode flippase, which excises the kanamycin resistance cassette, resulting in a complete gene deletion. pCP20 is temperature inducible (42°C), while pCMT-FLP is rhamnose inducible (final concentration 0.4%).

The resulting strains (BL21 *ΔrecA* and *W3110 ΔompT ΔrecA)* were transformed with pAVE011 plasmid, resulting in strains RA016 and RA015, respectively.

2.5 Chloramphenicol acetyltransferase assay

Chloramphenicol acetyltransferase assay (CAT assay) was performed to establish gene expression from pAVE011 and its mutated variants, using the modified "Enzymatic Assay of Chloramphenicol Acetyltransferase" protocol available online at: https://www.sigmaaldrich.com/GB/en/technical-documents/protocol/protein-biology/enzyme-activity-assays/enzymatic-assay-of-chloramphenicol-acetyltransferase.

Plasmids pAVE011CMR were isolated from overnight cultures inoculated with cryopreserved strain stocks from various experiments using the GenElute HP Plasmid

miniprep kit. The plasmids were then used as templates for PCR to confirm their promoter region length, from which their sequence was inferred. A fresh stock of competent DH5α cells was prepared and transformed with these pAVE011CMR plasmid variants. The freshly transformed cells were plated on LB agar plates (with tetracycline). The plates were incubated for a day at 37°C. Single colonies were picked from these plates, patched onto a square LB tetracycline (10 µg/ml) plate, and incubated at 37°C overnight.

The following reagents were prepared: 100 mM Tris buffer at pH 7.8, 2.5 mM 5,5'ditiobis(2-nitrobenzoic acid) (DTNB), 5 mM acetyl-CoA sodium salt and 0.3% chloramphenicol in purified water. Dissolution of chloramphenicol was facilitated by first dissolving it in 10% final volume of methanol.

Cultures of *E. coli* DH5α carrying pAVE011CMR plasmid variants were grown overnight at 37 °C. The overnight cultures were then used to inoculate fresh media in duplicate (at 1/20 dilution) and grown to the mid-log phase. One of the replicates was induced by adding IPTG to a final concentration of 0.5 mM. The cultures were then incubated for another 4 h.

The OD600 of cultures was measured, and 2 ml of each was centrifuged at 6000g for 10 min. The cell pellets were resuspended in 1 ml of 100 mM Tris buffer containing protease inhibitors (Pierce Protease Inhibitor Tablets). They were then lysed by sonication on ice (1 minute, 3 s pulse, 7 s rest time, 20% output). The protein concentration of the lysates was determined by the Bradford assay, according to the BioRad Bradford Assay Quick Start protocol. The readings were then compared to a standard curve prepared for the reader using bovine serum albumin standards.

A lysate volume containing 1 µg of protein was transferred into a 96-well plate, and the MilliQ water was then added to each well to make up the total volume to 10 µl

(except for lysates from cultures carrying recombined pAVE011CMR, where the volume transferred contained 10 µg of protein). Next, 26 ml of the 100 mM Tris buffer was mixed with 1 ml of DTNB solution, 1 ml of acetyl-CoA solution and 500 µl of the chloramphenicol solution. Finally, 190 µl of the mixed reagents was added to each well. Over several minutes, the absorbance at 412 nm was measured approximately every 0.5s.

2.6 6-well plate assay measuring fluorescent response in plasmid-carrying cultures.

Strains carrying pAVE011 or pIAH011 plasmids were investigated in response to varying IPTG concentrations and repeated exposure to IPTG (at a concentration of 0.5 mM).

The response to the inducer was recorded as the fluorescent response (in arbitrary units), measured in equal intervals (30, 15 or 10 minutes, depending on the experiment) alongside OD600 of the cultures on a 96-well plate reader (BMG FluoStar Omega). The plate in all experiments was incubated at 37°C, shaking at 200rpm.

In the repeated inducer exposure, a triplicate culture of one of the investigated strains was exposed to 0.5 mM IPTG in either the exponential or lag growth phase (referred to as pre-exposure). After overnight incubation, a sample of this exposed culture was taken to isolate the plasmid for PCR and gel electrophoresis of the promoter region. The rest of the sample was cryopreserved. In addition, the same treatment, except for IPTG addition, was applied to a separate triplicate culture of the same strain as negative (non-exposed) controls. The cryostocks of the pre-exposed and non-exposed cultures were then used to inoculate a 96-well plate, with each cryostock used to inoculate two wells. After those cultures reached OD600>0.4

following incubation, IPTG was added to one of the wells (referred to as “induced”), leaving the other well as negative control (“non-induced”). This procedure was repeated for all 4 of the strains.

The IPTG concentration-response experiment was also set up in a 96-well plate - a stock culture of one of the strains was used to inoculate all wells in a plate row. Then, it was induced after reaching OD600>0.4 with a serially diluted range of IPTG concentrations. This was repeated for all four of the strains.

2.7 Flow cytometry and cell sorting experiments

The cell sorting experiments were performed using the CytoFLEX SRT benchtop cell sorter and CytExpert software. Once a week, 100µl of the evolving cultures was transferred into fresh media in 24-well plates and grown at 37℃, shaking at 200 rpm for 1.5-2h. When the cultures reached OD600 of 0.3-0.5, IPTG was added to a final concentration of 0.5 mM to cultures (including controls). The cultures were left on the shaker to grow and express the fluorescent protein for several hours. Immediately before flow cytometry, the samples were diluted with fresh vLB media.

Using a fluorescent cell sorter, the top 10% of the fluorescent bacterial population was marked for selection from each of the six cultures of each of the four fluorescent strains (1515, 1516, IAH015, IAH016). Five thousand events (bacterial cells) were sorted from this high expression sub-population into new media on a 24-well plate. During sorting, the data on population heterogeneity was also collected.

The cell sorter lines were washed between samples with a solution of a detergent (FlowClean Cleaning Agent, Beckman Coulter), followed by deionised water. Whenever switching between strains, an additional washing step was performed with bleach (PureChloro10, 1% in deionised water) before detergent. The

wash solutions were run as a sample until no GFP signal was detected in the previously identified 10% most fluorescent subpopulation.

2.8 Data analysis methods

Most of the data analysis required some method of averaging to be used. Several mean averaging methods were used, depending on the distribution of experimental data. These are summarised in the **Supplementary Table S5**. Data transformation was also used to normalise the data distribution before performing some statistical tests.

The photographs of colonies on plates in the preliminary plate experiments were investigated using colour recognition software (Anny Studio, Just Color Picker) to classify them as green or non-green (protein-producing or not).

For the preliminary 96-well plate data organisation, the Microsoft Office Excel program was used. Once the data was arranged or the needed variables were calculated and extracted from the raw data, the analysis was performed in GraphPad Prism (versions 8.2.1 to 10.3.0, depending on the latest version available at the time of the analysis).

The experiment aiming to calculate the average promoter activity of the various strains post-induction involved some mathematical transformation of the raw data. An accurate comparison of the promoter activity between strains was possible due to applying an altered equation described in (Leveau, 2001). In this paper, the authors describe the following equation:

$$P = f_{ss} \times \mu \times (1 + \frac{\mu}{m})$$

Where:

P = Promoter activity (measured as a sum of transcription and translation)

f_{ss} = fluorescence steady-state constant described as the slope of FluorescencexOD600 plot

μ = bacteria growth rate described as the slope of the ln(OD600)x(time) plot

m = maturation rate constant calculated as ln(2)/(time constant of sfGFP maturation).

The time constant used to calculate m was 0.45h (Balleza et al. 2018)

In this study, the P value of calculated promoter activity is referred to as “estimated culture productivity”, which is more accurate in the context of this study.

3. Results I "Understanding the intrinsic instability of the pAVE011 vector under high expression pressure"

3.1 Introduction

Microbial fermentation is a well-established system for producing recombinant proteins. Usually, it involves transforming a bacterial host with a plasmid encoding the gene of interest, culturing the bacteria in a fermenter to reach maximum biomass, and inducing expression at the optimal time to maximise transcription and translation. One of the platforms used in this approach is pAVEway by FujiFilm, using *Escherichia coli* as the bacterial plasmid host (Lennon 2014).

PAVEway plasmids such as pAVE011 have a unique design **(Fig. 3.1)** featuring two palindromic *lac* operators on either side of the promoter. This allows for a DNA loop formation in the absence of an inducer via tetramerisation of DNA-bound LacI repressor protein dimers. This feature allows for precise control of the system. Under the control of the operator is a strong promoter, T7A3, originating from the T7 phage. Because it is a promoter controlling the expression of genes in early infection stages, it does not require phage polymerase (Mishra & Chatterji 1993). This is another advantage of the pAVEway system, as it does not require a genetically modified host (like BL21(DE3)), making it a versatile vector. The plasmid also encodes LacI repressor protein, which ensures enough repressor is produced to fully repress the system in the absence of an inducer.

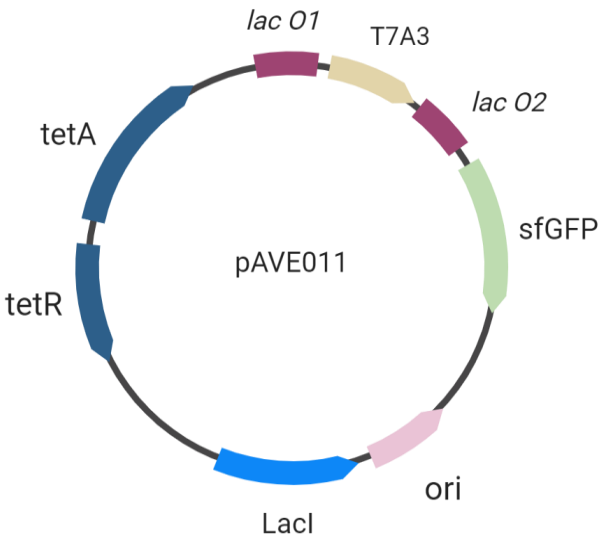


Fig. 3.1 The simplified plasmid map of the pAVE011 vector. The pAVE011 plasmid encodes tetracycline resistance genes (*tetR* and *tetA*), a *lac* repressor protein LacI, and a protein of interest (here, sfGFP). Two identical palindromic lac operator sequences surrounding the T7A3 promoter control the heterologous protein expression. T7A3 promoter is an early infection T7 phage promoter - strong and recognisable by the bacterial polymerase (Minkley & Pribnow 1973). Including the sfGFP coding sequence, the plasmid is 6454 bp long. It has a ColE1 *ori* and a medium copy number.

Additionally, to stabilise the pAVEway plasmids in the population, they encode tetracycline resistance genes *tetA* and *tetR*. TetA is a membrane-bound efflux protein, and its expression level is tightly controlled by the TetR repressor protein, as constitutive expression of *tetA* imposes a significant fitness cost to the host (Møller *et al.* 2016). The mechanism of tetracycline resistance is significant in the vector design, as it ensures that only cells that carry the *tetA*-encoding plasmid will express the efflux protein needed to confer the resistance. This is in contrast to other antibiotic-resistance genes, such as beta-lactamases, which are enzymes secreted into the environment which can protect antibiotic-susceptible clones in the population (Yurtsev *et al.* 2013). Finally, the pAVEway vector also contains the *cer* site, a resolvase target that prevents plasmid multimerisation and segregational instability (Summers & Sherratt 1988).

The goals of this chapter were to 1) characterise the pAVE011 vector stability in the host population, 2) introduce targeted modifications to improve the vector stability, 3) assess the resulting plasmids against the original ones in terms of productivity, inducer concentration required for production, stability, and other industrially relevant qualities, 4) obtain a host-plasmid combination that will be an improved starting point for co-evolutionary experiments.

Modifying the pAVE011 vector sequence could negatively impact the protein expression profile. Additional experiments comparing the new plasmid constructs against the original ones were designed to ensure that the modifications to the original vector do not result in unforeseen changes to protein yield, non-specific induction, and concentration of inducer required for production. Additionally, a second line of plasmids was created, carrying an inducible chloramphenicol gene on the pAVE011 plasmid instead of the initially encoded superfolder Green Fluorescent Protein (sfGFP). These plasmids were designed for replica plating experiments. All transformations were performed in the *E.coli* DH5α strain, which has a *recA1* point mutation disrupting the recombinogenic activity of RecA protein (Kawashima *et al.* 1984).

3.2 Results and Discussion

3.2.1 Imposing strong induction pressure is deleterious to protein production in pAVE011-carrying strains

To understand the experimental system provided by FDBK, we first undertook some simple preliminary experiments to examine the *E. coli* colonies when expressing GFP from the pAVE011-GFP plasmid. When grown on LB agar plates (with tetracycline) in the absence of any added inducer, the colonies were uniform and bright green, which was interpreted as “leaky” expression from the promoter present in the pAVE011 plasmid. However, when expression was induced by incorporating 0.5 mM IPTG into the agar, a mixed populations of small and large green colonies were seen together with pale or white colonies of various sizes in both *E. coli* genetic backgrounds (**Fig. 3.2A**). Imposing strong induction pressure on pAVE011-GFP carrying strains results in loss of sfGFP production.

In order to assess the genetic changes that might have led to the loss of superfolder GFP (sfGFP) production, colony PCR (and subsequent gel electrophoresis of the fragments) of the green and white colonies was undertaken. This used the pram1 primer pair (as described in methods) that amplified the promoter region. When used for PCR with the bright green-coloured colonies, it yielded the correct size band (479 bp) to indicate the promoter was intact **(Fig. 3.2B lanes 1-2)**. In contrast, when the same primer pair was used with the white colonies, the resulting products were shorter **(Fig. 3.2B, lanes 5-8)** and consistent with promoter deletion. In addition, in several clones, the population within a colony was mixed **(Fig. 3.2B, lanes 3, 4)**.

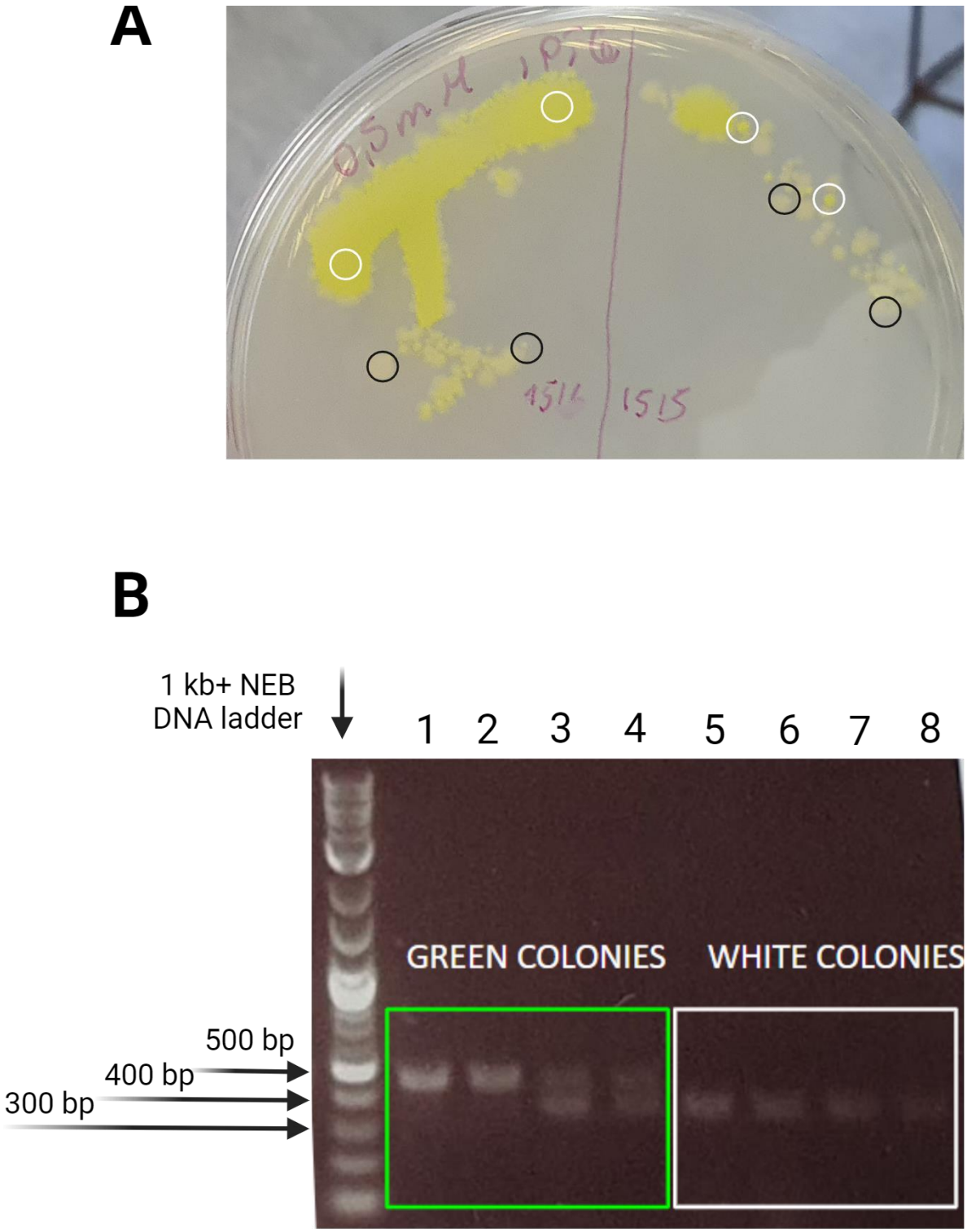


Fig. 3.2 Results of preliminary experiments. A) The photograph of 1515 and 1516 strains plated on LB agar plate containing 0.5mM IPTG. Encircled in white are colonies or areas where sfGFP is being overexpressed, and in black areas or colonies which are visibly paler in colour (lower sfGFP expression). The photographs were taken using a smartphone camera. The picture was deliberately left unaltered. Using a free image colour analysis software (Anny Studio, Just Color Picker), it was found

that the hue and value of the white and black encircled areas are similar. However, the saturation of the first is consistently above 50% and the latter below 50%). **B) Gel electrophoresis of the colony PCR products.** The primers were designed to amplify the promoter region. A predicted 91bp difference between the amplified region size depends on whether homologous recombination occurred. All white or pale colonies produced a shorter band consistent with promoter loss. All green colonies produced a longer band consistent with the intact promoter (in some cases alongside a shorter band, suggesting a mixed population).

Losing productivity due to inducer exposure in strains used for heterologous protein production is an undesirable trait. This can interfere with final product yield by introducing a non-productive subpopulation. It could also prove impossible to select for more productive clones during the future evolutionary experiments, as non-productive clones would have a fitness advantage over their productive counterparts. It was therefore important to understand what mechanism was causing the rapid productivity loss in pAVE011-GFP carrying strains exposed to IPTG on an agar plate.

It was hypothesised that the promoter deletion occurs due to homologous recombination between the two perfectly palindromic operators surrounding it. A schematic of the promoter region elements (original and recombined) is presented in **Fig. 3.3A** and 3.**3B,** respectively. The shorter bands identified due to homologous recombination from several experiments were sequenced **(Fig. 3.3C)**, confirming they correspond to the promoter and one of the palindromes being deleted. The preliminary results confirmed that the homologous recombination between operators occurs in pAVE011 plasmids, especially when bacteria are exposed to IPTG immediately upon inoculation. The resulting promoter deletion is likely a compensatory mutation alleviating the protein overexpression pressure. It has been previously reported that bacteria overexpressing sfGFP and its various protein fusions from a plasmid adapt to

the induction pressure by escaping expression via (among others) compensatory mutations in the promoter region, lowering its activity (James *et al.* 2021). Promoter deletion was also observed in pAVE011 plasmids carrying an inducible chloramphenicol resistance gene on agar plates containing this antibiotic. Sequencing data from 10 small-colony chloramphenicol-resistant clones carrying pAVE011CMR plasmid revealed recombination in the promoter region. To investigate the resistant phenotype of bacteria that have lost the original T7A3 promoter, the newly formed region upstream of the resistance gene was analysed using BacPP - a site designed for sigma factors binding sites prediction in *E. coli,* with overall accuracy above 80% (Silvaa *et al.* 2011). After homologous recombination in pAVE011 plasmid, a sequence arises **(Fig. 3.3C)** which is predicted to be recognised by σ38 and σ70 transcription factors (the main factors active in the expression of genes during the stationary phase of growth or stress, and exponential phase of culture growth (Tripathi *et al.* 2014) with a 95% probability. This, together with the smaller colony morphology of recombined chloramphenicol-resistant colonies compared to non-recombined ones (data not shown), indicates that after homologous recombination, the gene inserted into pAVE011 can still be expressed; however, from a weaker and no longer tightly controlled promoter. This finding rendered replica plating as an approach to estimating homologous recombination rate impossible, and an alternative method was designed.

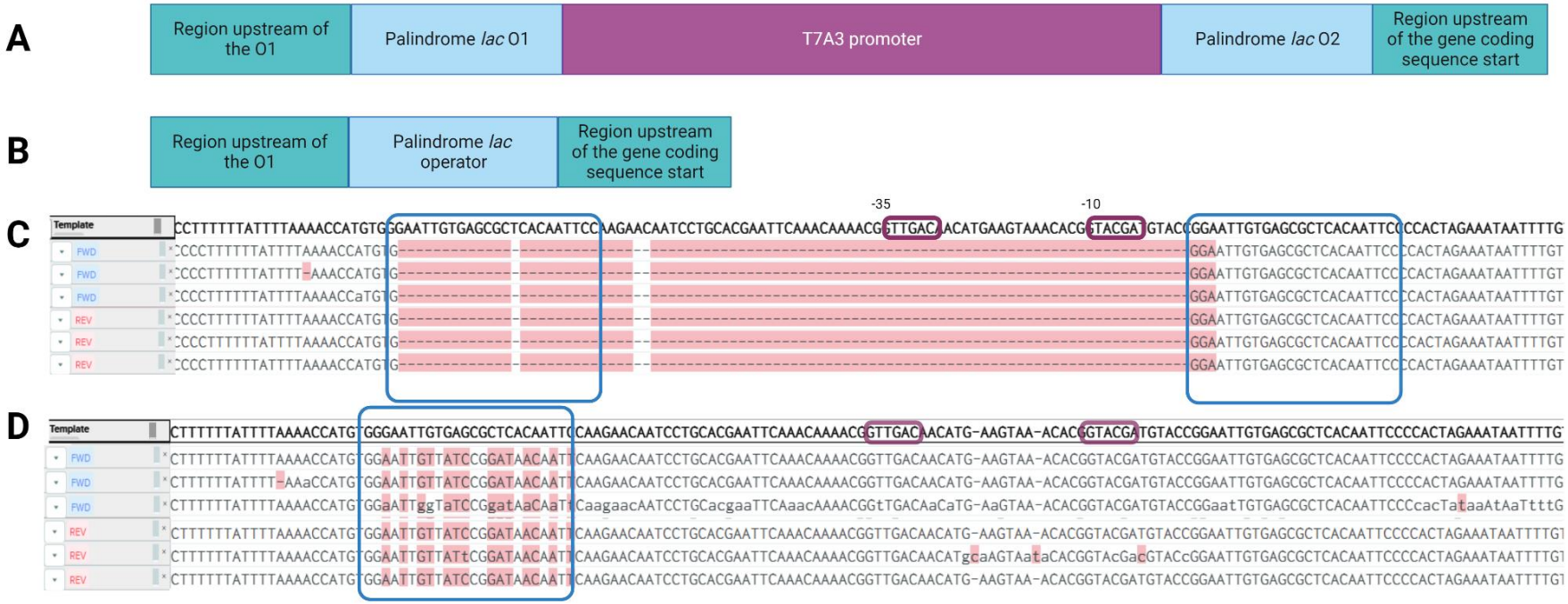


Fig. 3.3 Schematic representation of the sequences in the original and recombined promoter region of the pAVE011 plasmid

A) The pAVE011 promoter region diagram **B)** The pAVE011 promoter region following homologous recombination. This sequence of 147bp was analysed with BacPP, resulting in a high predicted probability (95%) of affinity to transcription factors σ70 an d σ38 in the 67bp to 147bp fragment. The sequencing results of the PCR amplified pAVE011 **(C)** and pIAH011 **(D)** promoter region from several independent experiments aligned to a pAVE011 template. Either a single colony or a boiled culture sample was used as the PCR template. Each forward (FWD) sequencing reaction is shown alongside a matched reverse reaction (REV). Eight pAVE011 and

five pIAH011 populations were analysed (three of each shown). Gel electrophoresis of the fragments (data not shown) was also performed, and the sequence predicted based on the band length was confirmed by these sequencing results. The red region in **C)** represents the deletion of the fragment. The blue boxes in **C)** and **D)** show the expected position of the operators, while the purple boxes show the position of the T7A3 -35 and -10 sequences.

3.2.2 Engineered palindromes reduce the frequency of homologous recombination in pAVEway plasmids

Homologous recombination between the two palindromic operators has been shown to occur in pAVE011 plasmids when exposed to the inducer (0.5 mM IPTG) in a plate environment (section 3.2.1). This reduces the plasmids' productivity and is a significant barrier to any following evolutionary experiments aiming to improve the relationship between the *E. coli* host and the plasmids. This is because the evolutionary pathways would likely involve recombination and thus result in plasmids with a lower expression rate and lesser value for industrial applications.

Palindromic or repeat sequences can be a hotspot for recombination, especially RecA-independent recombination, which relies on short regions of homology in close proximity (Chédin *et al.* 1994; Lovett *et al.* 1994; Mazin *et al.* 1991). Recombination in *E. coli* also relies on homology itself, which, when disrupted even by a single base pair in a 53 bp homology region, can decrease the recombination efficiency 4-fold (Watt *et al.* 1985).

To prevent homologous recombination in pAVE011 plasmids, an approach was designed to replace *the lacO1* operator (upstream of the promoter, **Fig. 3.4A**) with the alternative *lac O1'* sequence (**Fig. 3.4B**), creating pIAH011. This alternative palindromic sequence has a similar affinity to LacI repressor as the *lac O1* - 0.3kT difference in binding energy compared to 0.5kT-4.6kT for other *lac* operator sequences (Zuo & Stormo 2014). Therefore, the strict control over the pAVEway system (which relies on the DNA loop arising from a DNA-bound LacI tetramerisation) should not be altered; however, perfect homology between O1 and O2 DNA sites should be disrupted and prevent homologous recombination.

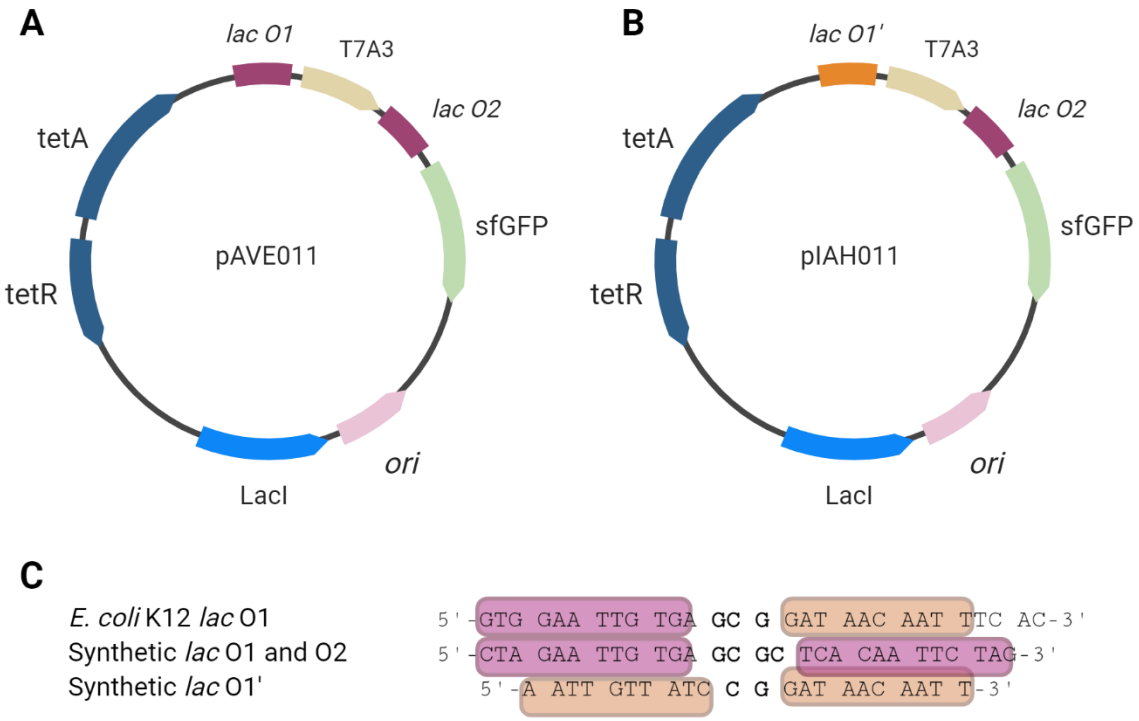


Fig. 3.4 The simplified plasmid maps of pAVE011 and pIAH011 and *E.coli lac* operator sequences. (A) The pAVE011 plasmid encodes tetracycline resistance genes (*tetR* and *tetA*), a *lac* repressor protein LacI, and a protein of interest (here, sfGFP). Two identical palindromic lac operator sequences surrounding the T7A3 promoter control the heterologous protein expression. T7A3 promoter is an early infection T7 phage promoter - strong and recognisable by the bacterial polymerase (Minkley & Pribnow 1973). **(B)** The pIAH011 plasmid map, resulting from replacing the *lac O1* sequence in pAVE011 with the *lac O1'* sequence. **(C)** The comparison of *E.coli lac* operator sequences. From top to bottom: native operator sequence in *E.coli lac* operon; perfect palindromic sequence present originally in pAVE011 plasmids; alternative palindromic sequence with similar affinity to LacI as the original palindrome.Homologous sequences are indicated with matching colours.

In order to evaluate whether this plasmid modification successfully reduced the recombination rate in the pAVE011 plasmid, 40 cultures of each of the four strains were incubated overnight at 37°C. These strains were original pAVE011-harbouring *E. coli* strains used by FujiFilm - 1515, a K background *E.coli* and 1516, a BL21 background *E. coli,* and isogenic strains carrying the modified pIAH011. FujiFilm uses these two genetic host backgrounds in fermentation, hence why they were tested. All of them were inoculated into LB containing 10 µg/ml tetracycline, and half also contained 0.5 mM IPTG. This immediate exposure to the inducer (in the lag phase of growth of the bacteria) mimicked the conditions of an agar plate, where high recombination rates were observed before (in contrast to low or no recombination when exposed in liquid culture in the exponential growth phase). The following day, boiled cultures were used as a template in PCR reactions, amplifying the promoter region of the plasmids. Boiled cultures were used as an alternative to colony PCR for several reasons. The experimental design included twenty biological replicates of each strain, which provided greater statistical power than testing several clones from fewer populations. This is evidenced by the results in **Fig. 3.5** and **3.6**, where most of the populations tested only contained either recombined or non-recombined promoter variant, not both. Secondly, picking clones from each culture would require transferring the cultures onto agar plates and allowing time for incubation. This would provide the clones with additional time for growth, allowing further mutations to occur. The PCR products were separated by gel electrophoresis **(Fig. 3.5 and 3.6)**. The expected sizes of the bands for recombined (lost) and intact promoters were 302 bp and 393 bp, respectively. An earlier preliminary experiment confirmed the expected sequence of the longer band **(Fig. 3.3D)**.

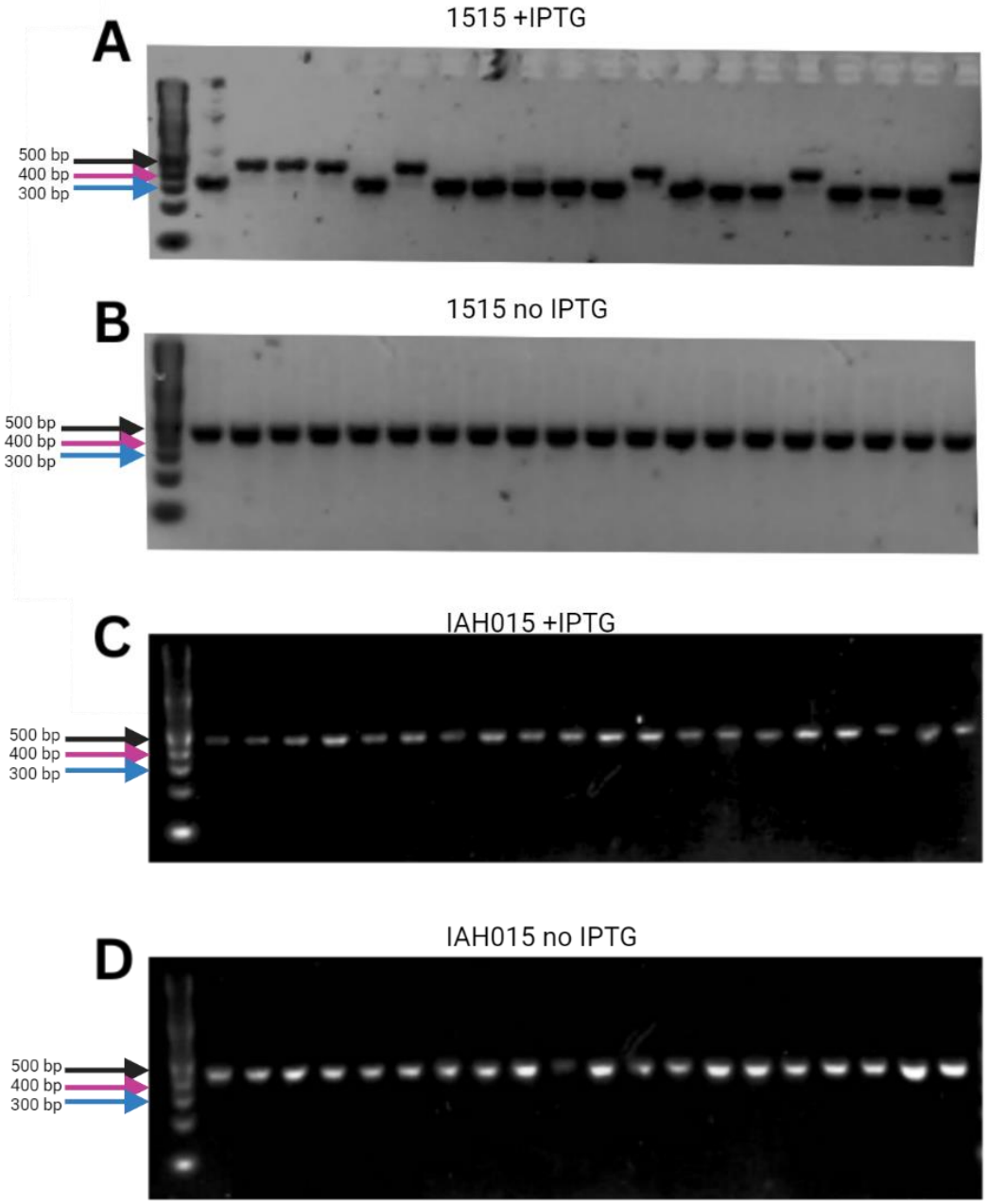


Fig. 3.5 The gel electrophoresis (3% agarose, TBE buffer, 50 V ran for 80 min) of promoter region PCR products from 1515 (A, B) and IAH015 (C, D) strains. All strains were grown overnight. Panels **A** and **C** show PCR products from the strains exposed to 0.5mM IPTG upon inoculation, and **B** and **D** show results from cultures grown without the inducer. Each lane represents one culture. The ladder used (NEB 1kb+) has been marked with coloured arrows at the following lengths: black at 500bp, magenta at 400 bp, and blue at 300 bp. The expected length of the intact promoter region is 479 bp, while that of the recombined promoter region is 388.

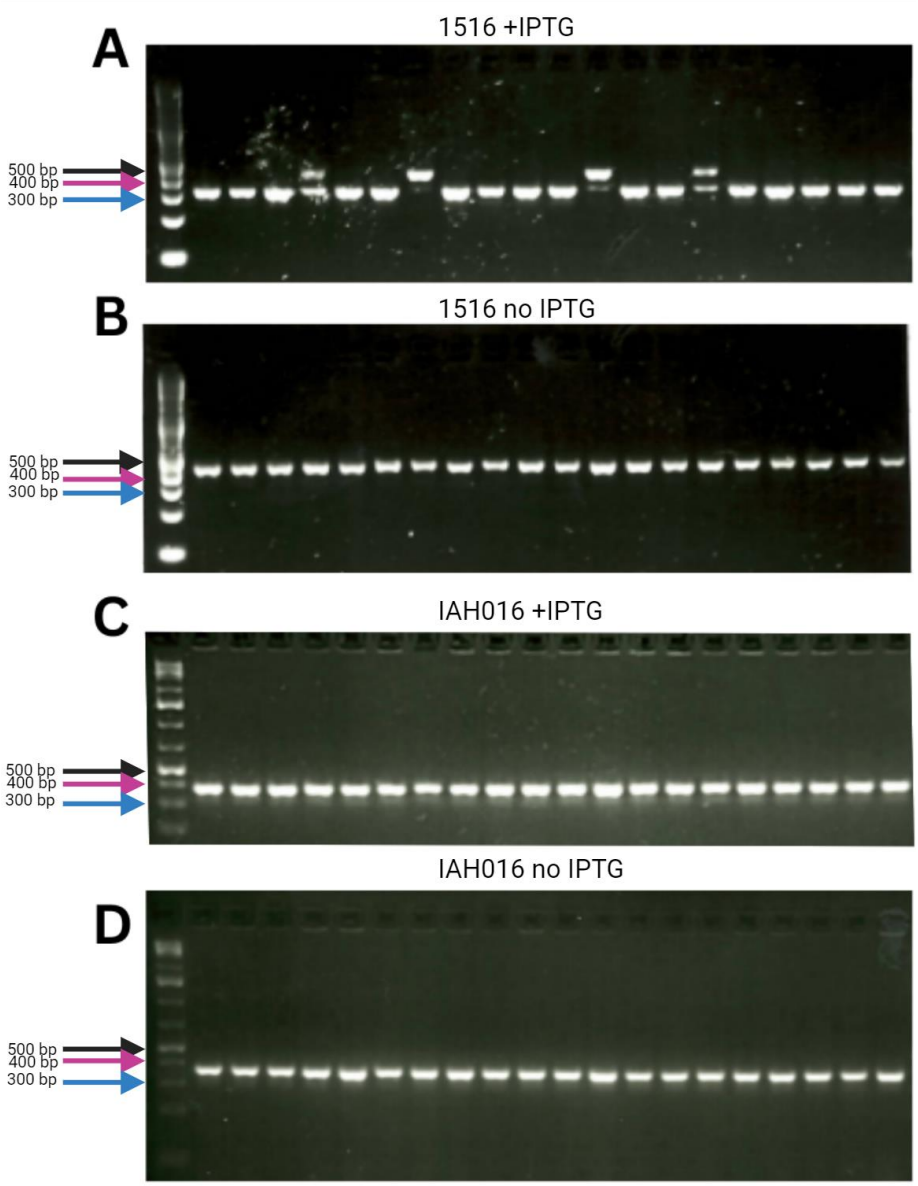


Fig. 3.6 The gel electrophoresis (3% agarose, TBE buffer, 50 V ran for 80 min) of promoter region PCR products from 1516 (A, B) and IAH016 (C, D) strains. All strains were grown overnight. Panels **A** and **C** show PCR products from the strains exposed to 0.5 mM IPTG upon inoculation, and **B** and **D** show results from cultures grown without the inducer. Each lane represents one culture. The ladder used (NEB 1kb+) has been marked with coloured arrows at the following lengths: black at 500 bp,

magenta at 400 bp, and blue at 300 bp. The expected length of the intact promoter region is 479 bp, while that of the recombined promoter region is 388.

The occurrence of recombination was counted **(Table 3.1)**, and results were analysed using a logistic regression approach. Triplicate results from an earlier experiment of a similar design were included in each group of the analysis in order to include extremely rare events in the model. In addition, mixed populations were counted as recombined since the model was supposed to estimate the homologous recombination event rate, and those did occur in mixed populations.

Strain	Exposure to 0.5mM IPTG in the lag phase of growth		No IPTG exposure	
	Operator recombination, promoter loss	No recombination, intact promoter	Operator recombination, promoter loss	No recombination, intact promoter
1515	16	7	1	22
IAH015	1	22	0	23
1516	19	1	0	20
IAH016	0	20	0	20

Table 3.1. Counts of recombination events in pAVE011- and pIAH011-carrying strains following exposure to 0.5mM IPTG immediately upon inoculation. The cultures of strains incubated overnight with or without an inducer were used as a PCR template to amplify the promoter region from the plasmids. The DNA fragments were separated using gel electrophoresis **(Fig. 3.5, 3.6)**, and based on their sizes, the cultures were classified as either recombined or non-recombined. Mixed cultures were classified as recombined. Data from an earlier triplicate experiment with a similar design was included in the analysis to establish a model estimating the rate of rare events. The above data was arranged in a multivariable table **(Supplementary Table S1)** and analysed using multiple logistic regression.

The logistic regression model is described below and the multivariate table constructed for analysis is available as **Supplementary Table S1.**

$$Y_{Odds} = \beta 0 \times (\beta 1^{X1}) \times (\beta 2^{X1 \times X2}) \times (\beta 3^{X1 \times X3})$$

Where:

Y_{Odds} describes the odds ratio of $\frac{recombination\ event\ occurring}{recombination\ not\ occuring}$

$\beta 0$ describes the model intercept (baseline recombination rate in K background strain harbouring pAVE011 (1515) not exposed to IPTG

$\beta 1^{X1}$ odds ratio multiplicative increase per change from no IPTG to 0.5mM IPTG in 1515 strain

$\beta 2^{X1 \times X2}$ odds ratio multiplicative increase per change from no IPTG to 0.5mM IPTG and from original to alternative operator (IAH015 strain)

$\beta 3^{X1 \times X3}$ odds ratio multiplicative increase per change from no IPTG to 0.5mM IPTG and from K background to B background (1516 strain)

The results **(Table 3.1, Fig. 3.5B)** show that the 1515 strain (K background with pAVE011 plasmid), which has not been exposed to IPTG at inoculation, has an odds ratio (OR) of homologous recombination occurring of 0.011 (95% CI 0.0006 to 0.0493, p<0.0001), which translates to a low recombination event probability of 1.09%. IPTG exposure **(Table 3.1, Fig. 3.5A)** raises this OR to 2.59 (95% CI 0.4382 to 50.543, p<0.0001), or a probability of 72.14%. However, this increase in homologous recombination odds after IPTG exposure was attenuated by alternative O1 in the plasmid **(strain IAH015; Table 3.1, Fig. 3.5C)** - OR 0.017 (95% CI 0.00093 to 0.0957, p<0.0001), or a probability of 1.67%, which was almost the same as that of a never-exposed 1515 strain. The host background (B vs K) had a non-significant effect on

homologous recombination rate in IPTG-exposed strains, with the OR for 1516 strain exposed to IPTG **(Table 3.1, Fig. 3.6A)** being 3.986 (95% CI 1.0979 to 15.42, p=0.5147), or probability of 79.92%. Together, the gel electrophoresis **(Fig. 3.5 and 3.6)** and the logistic regression results show that the *lac O1'* sequence introduced into the pAVEway plasmids successfully reduced the homologous recombination event rate in both genetic *E. coli* backgrounds (B and K) used by FujiFilm.

While it would have been beneficial to fit a more detailed model to investigate further the interactions between different factors and their influence on the rate of homologous recombination and the main effects of factors, this was not possible, as some events (such as recombination in pIAH011 plasmids not exposed to IPTG) were not observed; therefore the model could not be calculated due to perfect separation of the data points. Even so, the model predicts the recombination event rate well (AUC for ROC curve 0.9440 (95% CI 0.8995 to 0.9885, p<0.0001; Hosmer-Lemeshow hypothesis test 1.736, p=0.9423; log-likelihood ratio test 113.9, p<0.0001). However, it is worth noting that the model has a lower positive predictive power than the negative predictive power (76.09% and 98.55%). This means it will likely predict more homologous recombination events than observed in the experimental data. Designing an experiment with more replicate cultures could provide more data points, including occasional events, improving the model definition and fit. Nevertheless, this limited model is sufficient for this study, which evaluates whether the operator 1 change reduces the homologous recombination rate in pAVEway strains following immediate IPTG exposure.

3.2.3 Promoter deletion by recombination is *recA*-independent

Homologous recombination mechanisms in *E. coli* can be either RecA-dependent or RecA-independent. It has been previously reported that recombination between short repeats separated by a short non-homologous DNA sequence relies on the latter (Bi & Liu 1994; Chédin *et al.* 1994; Lovett *et al.* 1994). The homologous recombination between operators was observed not only in *E. coli* BL21 and W3110*ΔompT* but also in DH5α, a recombination-impaired strain containing the *recA1* variant of *the recA* gene (Song *et al.* 2015), which contains a single point mutation that disables its recombinase function in standard conditions (Bryant 1988). The mutant RecA1 protein cannot change conformation and bind to DNA and ATP.

Since *recA1* and *recA* sequences differ by a single base pair, it was considered that the *recA1* sequence reverted in the DH5α strain carrying recombined pAVE011 plasmid to the wild-type *recA* enabling the recombinase. To confirm the status of *recA* in 10 of the DH5α clones carrying recombinant pAVE011 plasmid, a ~500bp fragment of the *recA* region was amplified via PCR and sequenced, confirming the mutated *recA1* sequence present **(Fig. 3.7)**. These results suggest that *recA* independent recombination took place in this case. Therefore, *recA* gene knockout strains carrying pAVE011 plasmid have been designed to investigate the *recA* gene's influence on homologous recombination observed in pAVE011 plasmids further. The BL21 *ΔrecA* pAVE011 strain has been designated RA016, and W3110 *ΔrecA ΔompT* pAVE011 RA015.

Fig. 3.7 Sequencing results of a PCR amplified fragment of the *recA* gene from 10 DH5α colonies confirmed via plasmid promoter region PCR to carry the recombined variant of pAVE011 plasmids. Each PCR fragment from one colony was sequenced in two directions (forward - FWD and reverse - REV). The fragments were aligned using benchling.com to the DH5α RecA coding sequence, presented in the figure is a 163bp long fragment of the *recA* gene on the DH5α chromosome (starting at 2763795bp, ending at 2763957bp). The highlighted T is the single base pair mutation described previously to disrupt the encoded recombinase activity. The sequencing confirms that none of the recombined clones has reverted to the wild-type *recA* sequence.

The experimental design was similar to those investigating recombination rates in pAVE011- and pIAH011-carrying cultures. Here, 40 cultures of *ΔrecA* strains (RA016 and RA015) each were incubated overnight, and half of them were grown in media containing 0.5 mM IPTG. The following day, the promoter region in the cultures was investigated via PCR and the fragment gel electrophoresis **(Fig. 3.8)**. PCR fragments from overnight cultures of both B and K genetic backgrounds are just under 400 bp long, consistent with intact promoter structure (393 bp) and no recombination **(Fig. 3.8A, 3.8B)**. In contrast, the PCR fragments from overnight cultures grown in IPTG-supplemented media are shorter (close to 300 bp length, **Fig. 3.8 C, D)** and correspond with recombined promoter region length (302 bp). These results confirm that the recombination events observed in pAVE011-carrying *E.coli* strains are recA-independent.

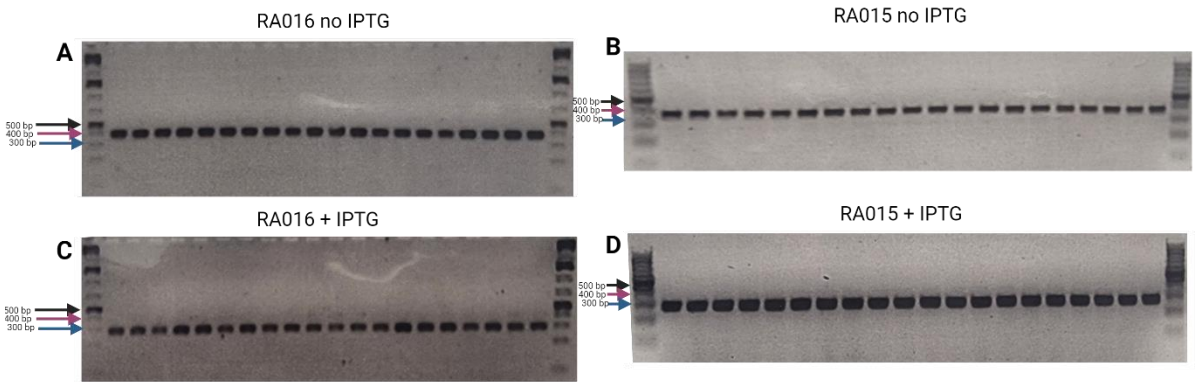


Fig. 3.8 Gel electrophoresis (3% agarose, TAE buffer, 35V, 80min) of promoter region PCR products from RA016 (A, C) and RA015 (B, D) strains. Strains were grown overnight. PCR reactions used templates from strains non-exposed to inducer **(A, B)** and exposed to 0.5 mM IPTG **(C, D).** GeneRuler 1kb+ ladder has been loaded in each gel's first and last well. The arrows point to the following fragment lengths: black - 500 bp; pink - 400 bp; blue - 300bp. The expected PCR product lengths are 393 bp and 302 bp for non-recombined and recombined promoter regions, respectively. Each sample lane represents one replicate culture.

Several RecA-independent homologous recombination mechanisms are involved in recombination between direct repeats or palindromes in *E. coli.* Secondary DNA structures forming within the palindrome are often cited as the cause of recombination, either through bringing the repeats separated by the palindrome closer together (Shukla & Roy 2006) or stalling the replication fork and leading to either DNA break repair mechanisms (Lovett *et al.* 1994) or DNA slippage (Lovett 2004). However, the recombination frequency dropping dramatically after replacing one of the palindromic operators in pAVE011 with a different palindrome makes it unlikely that a secondary structure forming within individual palindrome sequences stimulates the recombination. If that were the case, the alternative palindrome would stimulate the recombination at the same rate as the original. Instead, a structure involving both palindromes may be formed in the original pAVE011 plasmid. This relies on the palindrome homology, similar to mechanisms involved in recombination between DNA repeats. Usually, such a secondary structure results in a stalled replication fork resolved by sister strand exchange, and slippage during this process can lead to a plasmid dimer - of one original and one deleted sequence (Bzymek & Lovett 2001). In pAVEway plasmids, the dimers are resolved by the Xer-*cer* resolvase system (Hodgson *et al.* 2013), which is relevant in avoiding subsequent plasmid "dimer catastrophe" due to unresolved dimers accumulation (Field & Summers 2011). The altered operator sequence in the pIAH011 plasmid makes forming the between-palindrome secondary DNA structure impossible due to disrupted homology, reducing the rate of homologous recombination. It is unknown which *recA*-independent, short repeat homology-reliant mechanisms play a part in recombination events in pAVE011 plasmids. Although it could be investigated further via known recombination gene knockouts, it lies beyond the scope of this study.

3.2.4 Evidence of promoter activity following plasmid promoter deletion via homologous recombination

The plasmid sequence arising from promoter deletion by recombination in pAVE011 strains has been presented earlier in this chapter (**Fig 3.3B, C**), together with computational evidence from a bacterial promoter prediction program (BacPP) suggesting that bacterial transcription factors may recognise this new sequence. Earlier experiments have also demonstrated that 1) exposure to IPTG in vulnerable growth phases leads to plasmid recombination and 2) regardless of the exposure status, the pAVE011CMR-carrying strains maintain their chloramphenicol resistance. While the activity of this hypothetical promoter was evident from the resistant phenotype, the levels of this activity were not known. Therefore, it was essential to establish differences in promoter activity levels between strains carrying recombined and non-recombined plasmid.

Initially, growth assays were performed to establish minimum inhibitory concentrations (MIC) for recombined and non-recombined plasmid-carrying clones. These attempts were unsuccessful and required an alternative approach.

Twelve clones of DH5α pAVE011CMR were isolated from cryopreserved stocks, which were either previously exposed to IPTG or never exposed to IPTG (six of each). These clones were then used as colony PCR templates to amplify both the plasmid promoter region and a 600bp *recA1* gene coding sequence fragment encompassing the point mutation responsible for the inactivation of the recombinase product. The plasmid fragment sequence was inferred from fragment length, while the *recA1* fragments were sequenced using Sanger sequencing. The promoter region fragment provided information about the plasmid recombination status, while the *recA1* fragment sequence was analysed to confirm the inactivity of its protein product.

The gel electrophoresis of the fragments confirmed that clones previously exposed to IPTG carried recombinant plasmids with promoter deletion **(Fig. 3.9, lanes 16-21)**, while those never exposed to IPTG carried intact plasmids **(Fig. 3.9, lanes 2-7)**. At the same time, *recA1* PCR products have been sequenced. Each fragment's forward and reverse reaction resulted in two sequence reads per clone, which were then assembled in Geneious Prime to produce a consensus. These consensus sequences have then been aligned to the *recA1* reference sequence from reference *E. coli* genomes available online (Accession: AP009048; CP010816; CP017100)

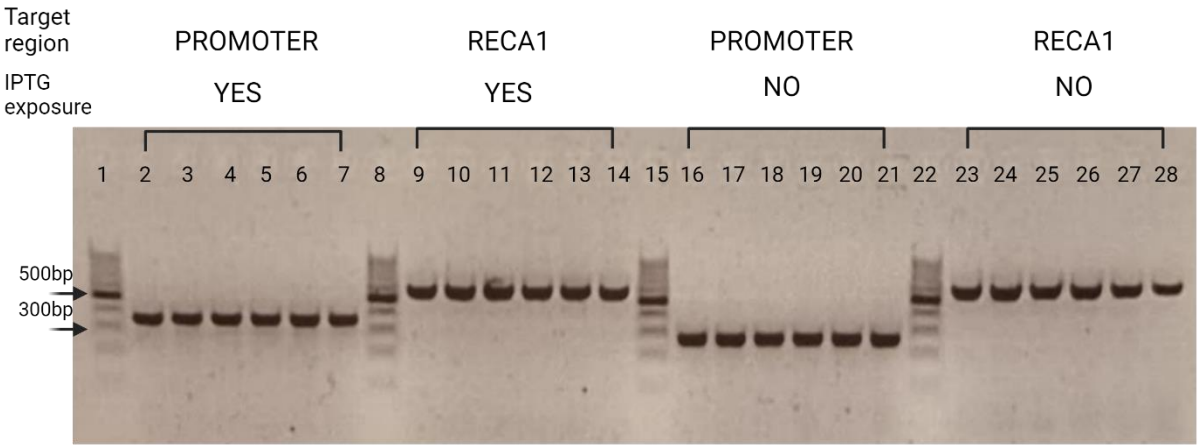


Fig. 3.9 The gel electrophoresis (3% agarose, TAE buffer, 100V, 40min) of PCR amplified plasmid and genome fragments originating from DH5α pAVE011CMR strains. Lanes **1, 8, 15, 22** - ThermoFisher GeneRuler 100bp; lanes **2-7** - amplified plasmid promoter region from 6 DH5α pAVE01CMR colonies never exposed to IPTG; lanes **9-14** - amplified *recA1* fragment from the same colonies; lanes **16-21** - amplified plasmid promoter region from 6 DH5α pAVE011CMR colonies previously exposed to IPTG; lanes **23-28** - amplified *RecA1* fragment from the same colonies. All strains have been confirmed as chloramphenicol-resistant. Each clone was used in two PCR reactions to amplify the plasmid promoter region and a fragment of *recA1* encompassing the location of point mutation inactivating the gene. The expected fragment sizes were 600 bp *recA1* coding region fragment, 360 bp intact plasmid promoter region, and 269 bp recombined plasmid promoter region.

The *recA1* gene variant in DH5α carries a single point mutation (**Fig. 3.10A**, black arrow), which results in the glycine residue of functional protein being replaced by aspartic acid. Neither clones exposed to IPTG **(Fig. 3.10B)** nor not **(Fig. 3.10C)** have been found to revert to the original sequence, which would have restored *recA1* functionality. This further supports the conclusion that recombination in pAVE011 plasmids is a *recA*-independent process.

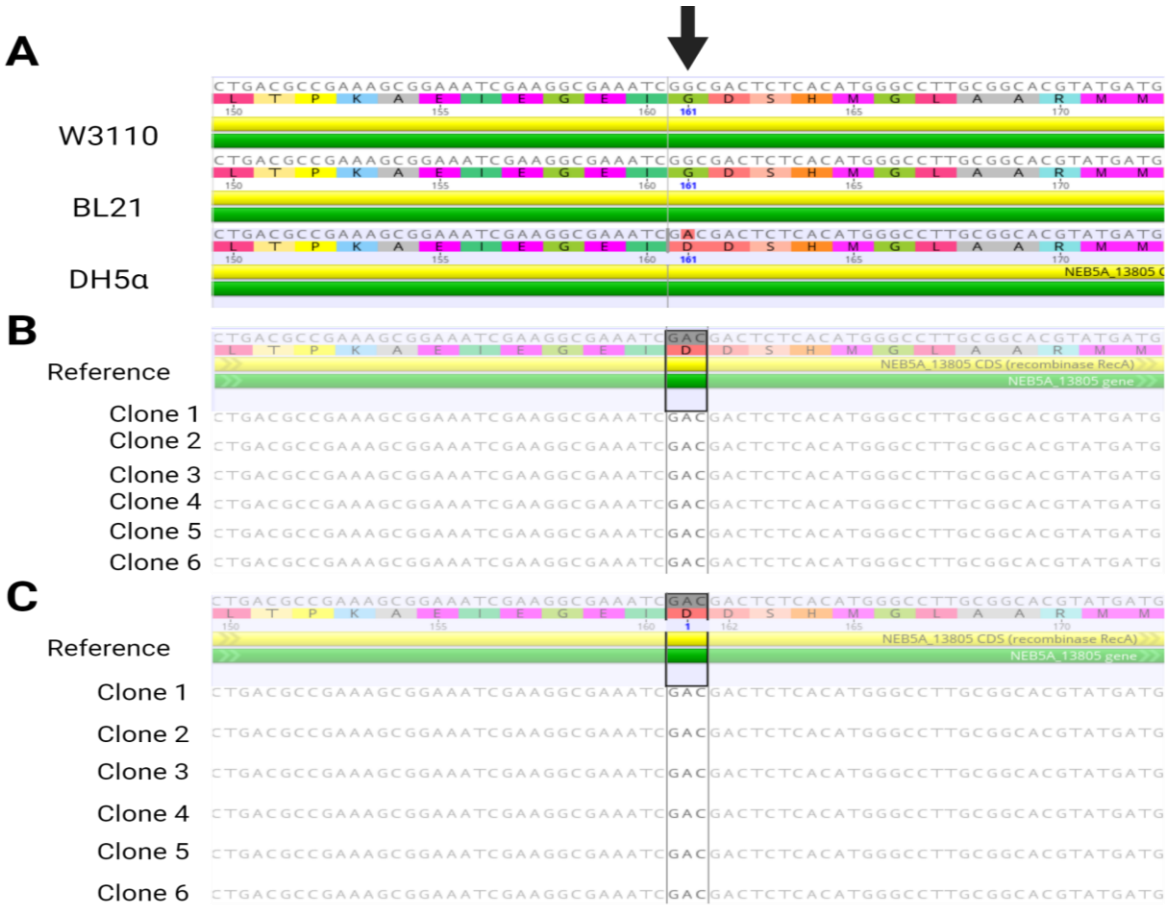


Fig. 3.10 Gene sequence alignments of *recA* coding sequence region. A) The alignment of *recA* regions from two strains with functional recombinase (W3110 and BL21) with DH5α. The black arrow points to the mutation, resulting in the *recA1* gene variant. The same region from DH5α was used as a reference sequence, to which 12 consensus sequences from each forward and reverse sequencing reactions of 12 DH5α clones were aligned. **B)** 6 clones were chloramphenicol resistant, never exposed to IPTG and carried intact pAVE011 plasmid. **C)** 6 clones were chloramphenicol resistant, previously exposed to IPTG and carried recombined pAVE011 plasmid.

After confirming the genotype of the plasmids (recombined and non-recombined) present in 12 chloramphenicol DH5α pAVE011CMR clones, it was necessary to confirm that the resistant phenotype was related to plasmid carriage and not changes in the bacterial genome. To this end, both recombined and non-recombined plasmids were isolated using a commercial plasmid miniprep kit and transformed into plasmid-free DH5α host from a laboratory -70°C stock as described in the methods.

While the resistant phenotype was known and confirmed to be directly related to pAVE011CMR plasmid carriage (both recombined and non-recombined) it was impossible to compare the levels of expression from the original pAVE011 promoter and the hypothetical one forming after homologous recombination in pAVE011 plasmids. Directly measuring the chloramphenicol acetyl-transferase activity in cellular lysates from cultures carrying recombined and non-recombined pAVE011-CMR was required in order to make that comparison between two promoter activity levels.

Twenty-four cultures of *E. coli* DH5α were grown overnight in LB: 6 carried empty pAVE011 plasmid; 6 carried pAVE011CMR; 6 carried pAVE011CMR with recombined promoter region and 6 carried no plasmid. The overnight cultures were then used to inoculate fresh media in duplicate (at 1/20 dilution) and grown to the mid-log phase. One of the replicates was induced by adding IPTG to a final concentration of 0.5 mM. The cultures were then incubated for another 4 h.

The OD600 of cultures was measured, and 2 ml of each was centrifuged at 6000g for 10 min. The cell pellets were resuspended in 1 ml of 100 mM Tris buffer containing protease inhibitors (Pierce Protease Inhibitor Tablets). They were then lysed by sonication on ice (1 minute, 3 s pulse, 7 s rest time, 20% output). The protein concentration of the lysates was determined by the Bradford assay, according to the

BioRad Bradford Assay Quick Start protocol. The readings were then compared to a standard curve prepared for the reader using bovine serum albumin standards **(Fig. 3.11).**

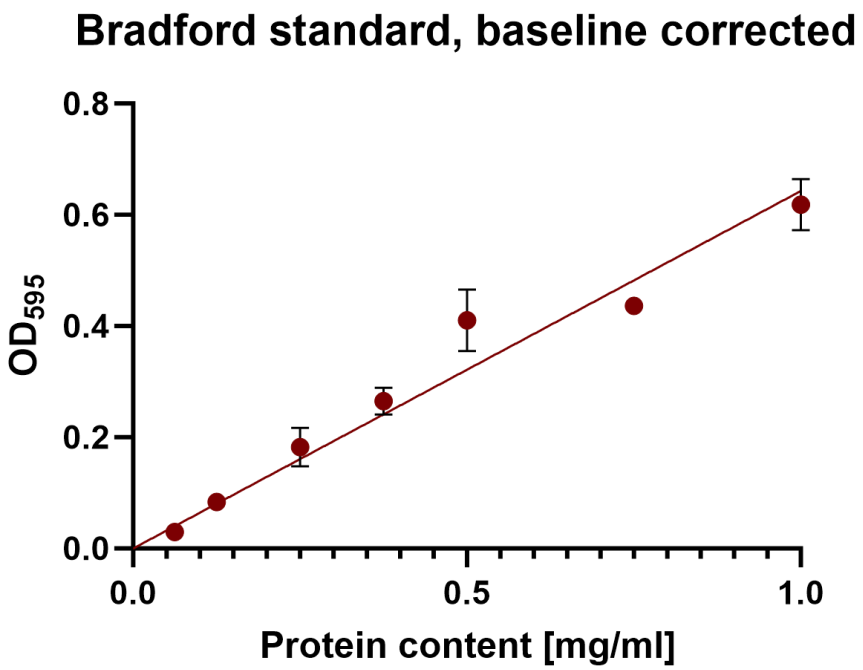


Fig. 3.11 The Bradford standard curve used to determine protein concentrations in the lysed samples. The line was fitted using GraphPad Prism 10.0.2 simple linear regression function. The slope was determined as 0.6428 with 95% CI 0.6135 to 0.6720.

With the resulting values of protein content per millilitre of lysate, the collected values of total culture volume harvested, and total OD1 equivalent harvested, it was possible to calculate the total protein per OD unit present in the induced and uninduced cultures **(Fig. 3.12).** There were no statistically significant differences observed between induced and uninduced cultures and their total protein content per OD600 unit (2-way ANOVA).

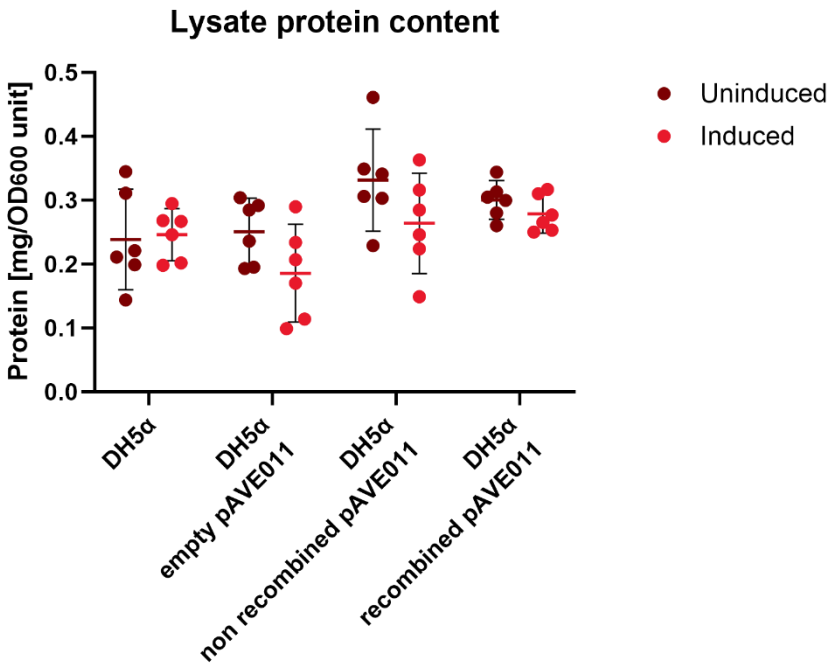


Fig. 3.12 The calculated total protein content per OD600 unit harvested in pAVE011 carrying culture lysates. The DH5α carries no plasmid and is a negative control. Similarly, the empty pAVE011 lacks the recombinant protein coding sequence. Each data point represents a single lysate. The arithmetic means, and their 95% CI bars are represented by the bars.

The above data is insufficient to conclude the promoter strength in recombined and non-recombined pAVE011 plasmids, as it only assesses the overall protein content per OD600 unit of each culture. The differences between induced and uninduced culture protein concentration cannot be attributed to chloramphenicol acetyl-transferase encoded by the plasmid alone, as the Bradford assay is non-specific. Therefore, a chloramphenicol acetyl-transferase assay (CAT) was performed using the cell lysates to act as enzyme solutions. The assay relies on two subsequent reactions. In the first one, Acetyl Coenzyme A (Acetyl-CoA) reacts with chloramphenicol (reaction facilitated by chloramphenicol acetyltransferase encoded on the plasmids), producing Coenzyme A (CoA) byproduct. The second reaction is the spontaneous reaction of CoA with DTNB (5,5'-Dithio-bis (2-Nitrobenzoic Acid)),

producing 5-Thio-2-Nitrobenzoic Acid (TNB). TNB can be detected by measuring OD412, and the change in absorbance can be used to determine Units/ml of enzyme solution added to the reaction. Here, the enzyme solution is the bacterial lysate. The Units of enzyme can then be expressed as Units per 1 mg of total cell protein and Units per OD600 harvested.

The collected absorbance 412 readings from the control reactions have been plotted over time **(Fig. 3.13).** A straight line has been fitted to all data points and plotted, and the slopes of these lines have been recorded **(Fig. 3.14).**

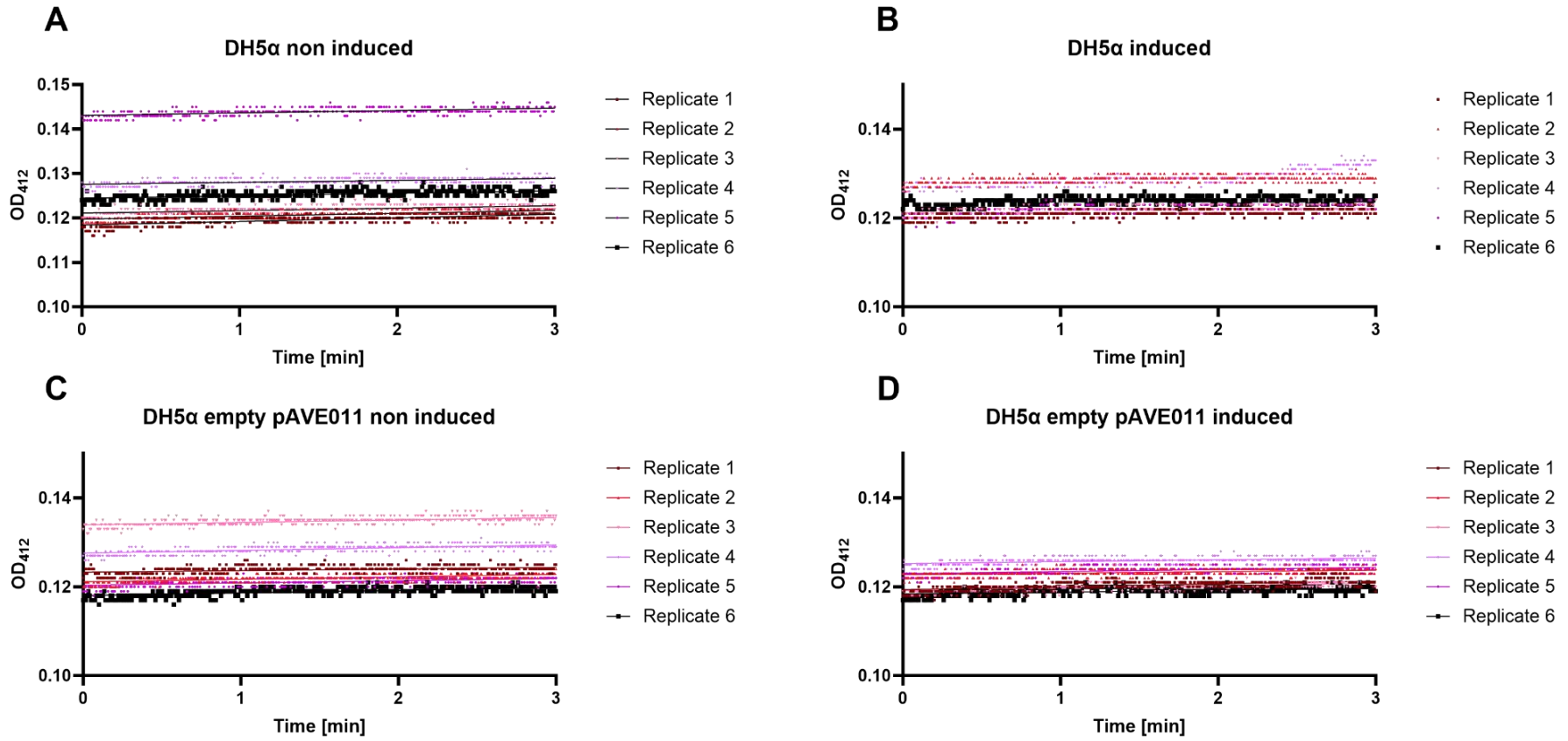


Fig. 3.13 Chloramphenicol acetyl-transferase assay negative controls. As negative controls for the assay, the untransformed DH5α ***(A, B)*** and DH5α transformed with pAVE011 with no heterologous protein coding sequence ***(C, D)*** were used. A lysate was obtained from each strain, either uninduced ***(A, C)*** or induced ***(B, D)***. These lysates were then used in the assay, and an absorbance reading (412nm) was taken every 0.5 seconds. The linear regression was performed using GraphPad Prism (10.0.2)

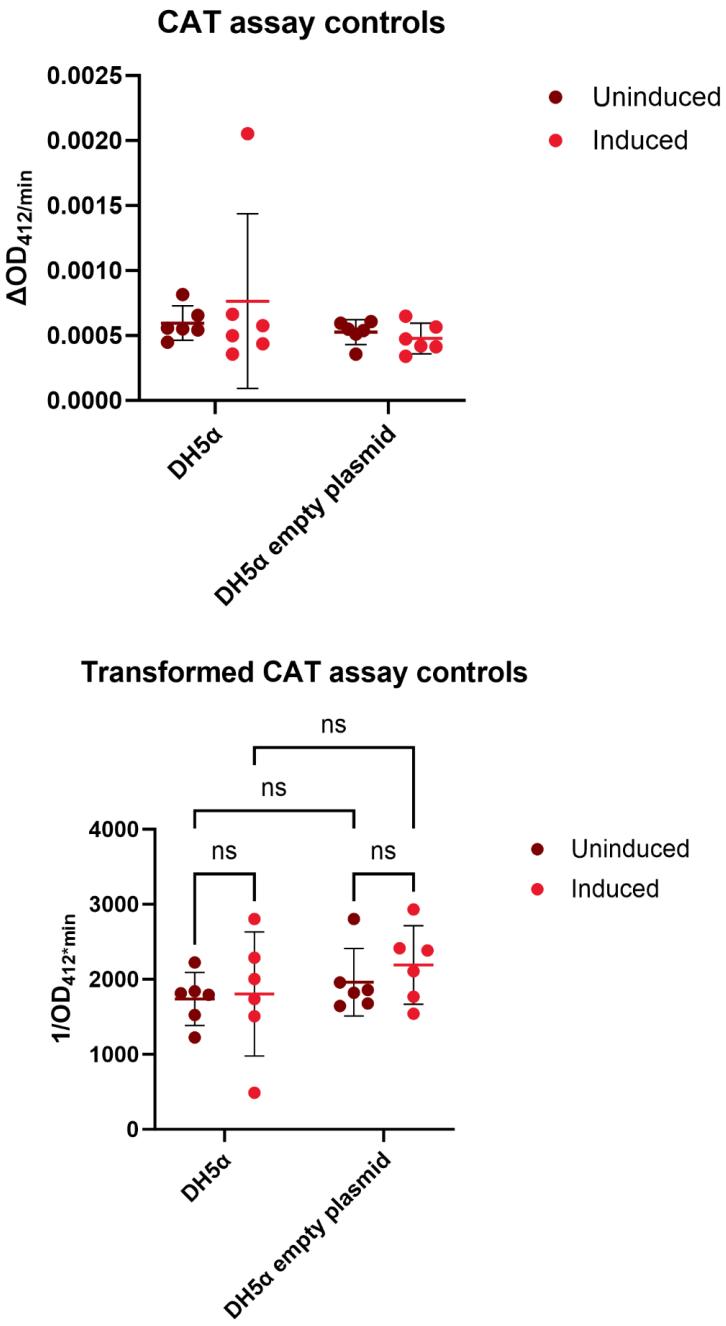


Fig. 3.14 Chloramphenicol acetyl-transferase assay negative controls - changes in Absorbance (412nm) per minute. The bars represent arithmetic means with its 95% CI. In order to normalise data distribution before statistical analysis, the data was transformed (1/Y). The statistical significance relates to Uncorrected Fisher's LSD test run after 2-way ANOVA.

As expected, the average changes in absorbance (412nm) per minute in all four negative control reactions were very low (less than 0.001/min). Neither induction nor empty plasmid carriage affected the calculated absorbance change rate. Therefore, all values were pooled, and the harmonic mean of 0.0005197 was used in further calculations.

The changes in absorbance over time were also plotted for the lysates of pAVE011CMR carrying cultures **(Fig. 3.15).** The rate of absorbance change was

calculated from the earliest linear segment of each plot, fitted in GraphPad Prism 10.0.2. All lysates were diluted so that approximately 1 µg of protein was added to each assay reaction (except for lysates of cultures carrying recombined pAVE011CMR, where protein content has been normalised to 10 µg). The increased amount of lysate added to the assay for these cultures was crucial to collecting the absorbance change data from the true linear part of the graph.

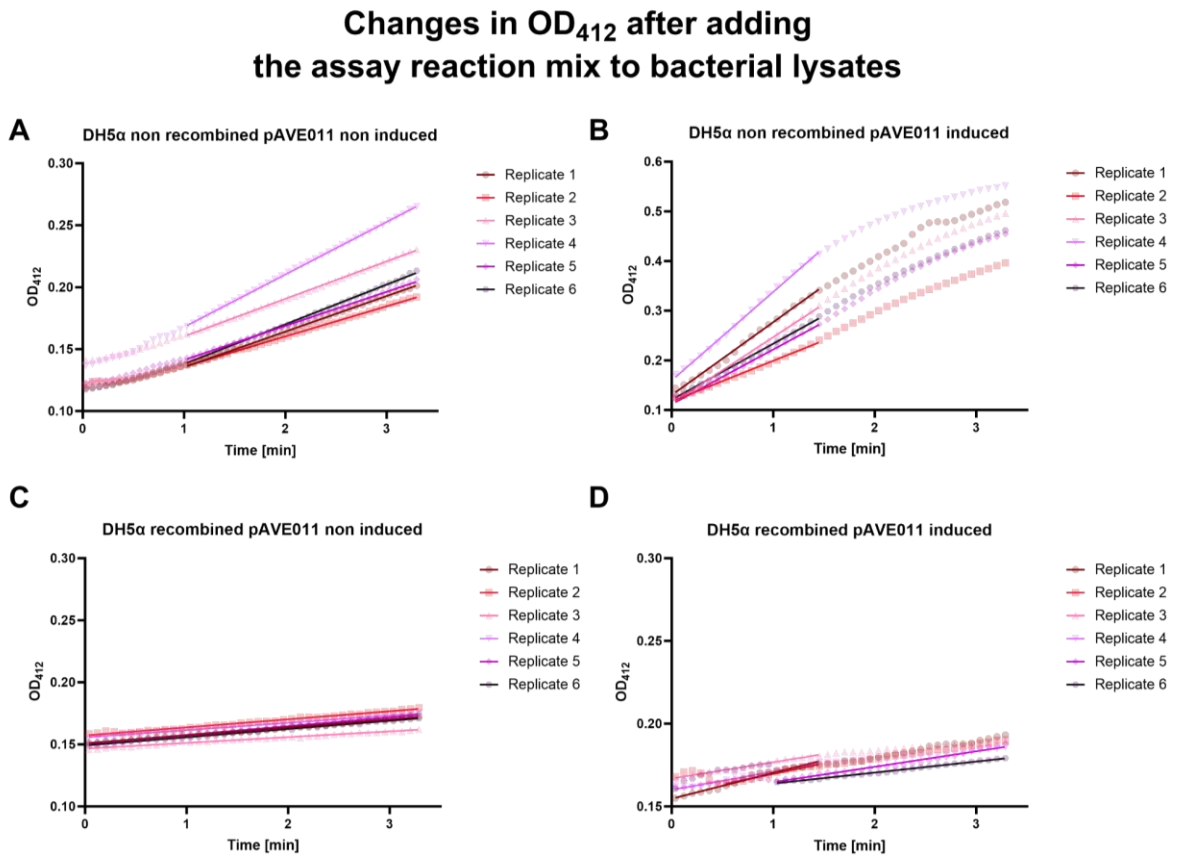


Fig. 3.15 CAT assay absorbance readings for lysates of DH5α carrying non-recombined (A, B) and recombined (C, D) pAVE011CMR. The cultures were either non-induced **(A, C)** or induced **(B, D).** All collected data until 2 minutes after lysate addition into the reaction mix is plotted. The lines represent lines of best fit in the initial, linear part of the charts.

Lysates of all cultures showed a steady, approximately linear (in most cases) increase in absorbance over time **(Fig. 3.15)**. The earliest linear graph segment differed between lysate groups. The rate of increase in absorbance was higher in the induced samples (carrying either variant of the pAVE011CMR plasmid, **Fig. 3.15 B, D**), suggesting more enzyme present in the reaction. The lysates from non-recombined induced pAVE011CMR carrying cultures caused an immediate sharp increase in absorbance within the first several readings **(Fig. 3.15 B).**

The changes in absorbance over time alone are insufficient to estimate the enzyme content of the lysates. To do that, the following calculations were done:

$$\frac{Units\ of\ enzyme}{ml\ of\ lysate} = \frac{(\Delta A412nm/min\ Test\ -\ \Delta A412nm/min\ Control) \times\ AV\ ml\ \times\ (DF)}{0.0136 \times LV}$$

Where:

Test - the reaction for which the units of enzyme/ml of lysate are being calculated

Control - the negative control reaction (or their average); here, constant 0.0005197

AV - assay volume, constant 0.2 ml

DF - dilution factor of the lysate

0.0136 - micromolar extinction coefficient of TNB at 412 nm

LV - lysate volume, constant 0.01 ml

$$\frac{Units\ of\ enzyme}{mg\ of\ protein} = \frac{Units/ml\ enzyme}{mg\ protein/ml\ lysate}$$

$$\frac{Units\ of\ enzyme}{OD600\ unit\ of\ culture\ harvested} = Units\ of\ enzyme/mg\ protein\ \times\ mg\ protein/total\ OD600\ harvested$$

The equations were adapted from an online protocol (Merck KGaA n.d.).

The calculated enzyme units per OD600 unit of culture harvested have been plotted **(Fig. 3.16)** and compared using two-way ANOVA with Uncorrected Fisher's LSD on log-transformed data. The transformation was necessary as ANOVA is a test comparing differences between arithmetic means. In this case, comparing how many times one mean was bigger than the other (differences between geometric means) was more beneficial. The results show that the overall yield (units of enzyme per OD600 unit of induced culture) was 132 times higher in non-recombined cultures than in recombined cultures (Uncorrected Fisher's LSD $p<0.0001$; 95% CI 80.7 to 217.3 DF=20; t=20.59). However, the non-recombined pAVE011CMR culture lysates also contained 49 times more enzyme per OD600 unit harvested than their recombined counterparts in the absence of inducer (Uncorrected Fisher's LSD $p<0.0001$; 95% CI 30.1 to 80.9 DF=20; t=16.43). Finally, the non-recombined pAVE011CMR carrying strains show 5-fold increase of enzyme units per OD600 culture harvested upon induction (Uncorrected Fisher's LSD $p<0.0001$; 95% CI 3.6 to 7.8 DF=10; t=9.466), while the recombined pAVE011CMR carrying strains show only a 2-fold increase (Uncorrected Fisher's LSD $p=0.0033$; 95% CI 1.3 to 2.9 DF=10; t=3.84). Together, this data confirms the presence of a new promoter-like sequence in the recombined region of the plasmid, which is controlled by the remaining *lac* operator sequence. It also suggests that the original promoter sequence is at least 2.7 times stronger than the recombined sequence.

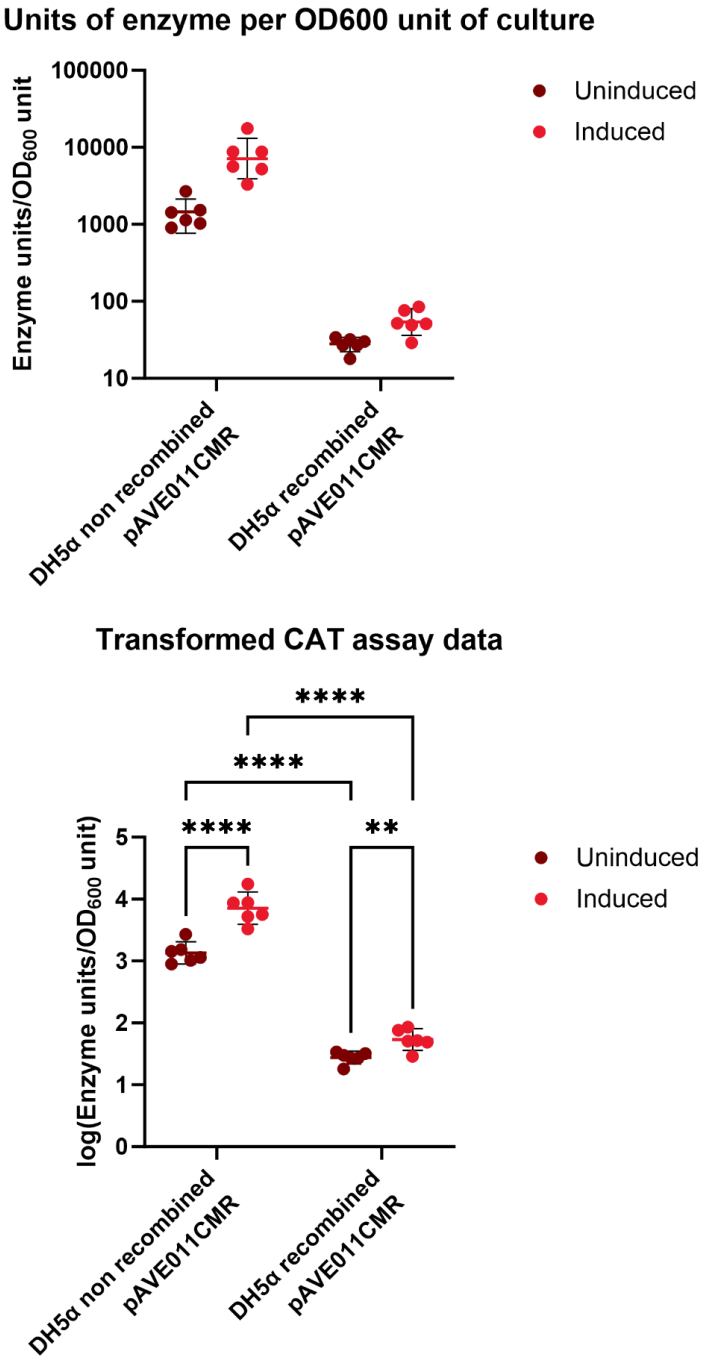


Fig. 3.16 The chloramphenicol acetyltransferase content per OD600 unit of culture collected. The DH5α cultures carried either recombined or non-recombined pAVE011CMR. The cells were harvested following subculturing into fresh LB media and induction. The enzyme units were calculated in two steps: 1. Calculate enzyme units per mg of protein present in lysate from units of enzyme per ml of lysate and mg of protein per ml of lysate; 2. Calculate the units of enzyme present in each OD unit harvested from the units of enzyme present per mg protein in lysate and mg of protein harvested per each OD unit. The statistical analysis (Two-way ANOVA with Uncorrected Fisher's LSD) was performed on log-transformed data. Asterisks refer to the following p-values: p≤0.05 - *; p≤0.01 - **; p≤0.001 - ***; p≤0.0001 - ****.

3.2.5 Method for accurate quantification of differences between pAVE011 and pIAH011 plasmids performance

After establishing that the substitution of the O1 sequence in pAVE011 plasmid (resulting in the pIAH011 variant) has reduced the rate of homologous recombination, it was essential to investigate whether this alteration had a significant effect on protein production and strain performance in a mock induction protocol based on the protocol used in fermenters. While the promoter sequence of the plasmid was not changed, replacing the O1 sequence with O1' sequence could have caused several problems with relevant implications for the fermentation, for example:

- changes in LacI tetramer formation on the DNA, and in consequence, loss of the fine-tuning control of expression in the system
- altered requirements for inducer concentration needed for induction (which may affect the cost of production)
- reduced growth rate related to the host's inability to alleviate the protein expression pressure via homologous recombination and promoter loss (affecting the time needed to achieve the maximum biomass)
- new evolutionary processes in the promoter region altering its affinity to polymerase

The aim was to establish the differences in 1) total sfGFP accumulation in cultures harbouring pAVE011 or pIAH011 in 20h post-induction period, 2) promoter activity defined as the sum of transcription and translation rates, 3) growth rate of the hosts, 4) all of the above host parameters in response to repeated IPTG exposure. To this end, an experiment was designed (fully described in **Fig. 3.17**), which consisted of pre-exposing the bacteria to 0.5 mM IPTG in one of the two exposure protocols (lag growth phase, immediately upon inoculation or exponential growth phase) and

subsequent 96-well plate assay where those pre-exposed strains (and relevant controls) were induced (or not) after OD600>0.4. The fluorescence and growth data were collected every 30 minutes during the 24-hour incubation.

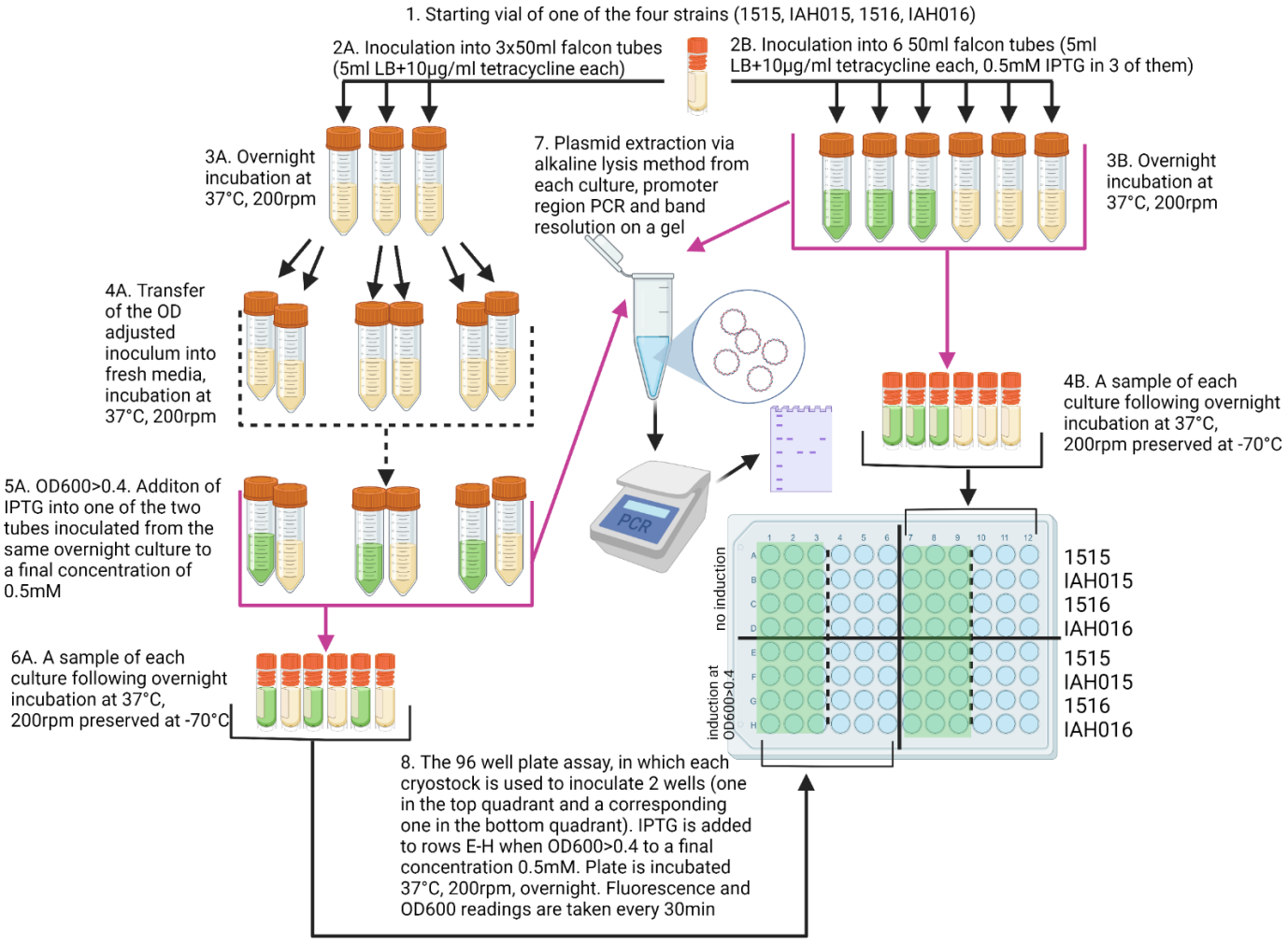


Fig. 3.17 A plate assay design for investigating pAVE011 (1515, 1516) and pIAH011 (IAH015, IAH016) carrying strains' performance under induction stress. Two protocols were designed to probe the influence of the first IPTG exposure time (either in the exponential growth phase of the culture, A, or lag phase, B) on protein production and growth. **1.** A single vial of one of the strains is used to inoculate overnight cultures **2.** The media contains no IPTG **(A)** or 0.5mM IPTG in half of the tubes **(B) 3.** Cultures were incubated overnight **(A and B) 4A.** OD600 of cultures was measured, and they were diluted with fresh LB to a standard OD600 value (0.8). These three cultures were immediately used to inoculate fresh media in two tubes each. These are grown for approximately three hours until OD600>0.4. **5A.** IPTG is added to one of the two cultures originating from one overnight culture to a final concentration of 0.5 mM. All six cultures are returned to the shaker and

incubated at 37°C overnight. **6A and 4B** A sample of each culture is mixed with sterile 50% glycerol stock and stored at -70°C. **7.** A sample of each culture is used to isolate plasmids, which are then used as templates in PCR reactions amplifying the promoter region. The fragments were resolved on an agarose gel, and a conclusion was made about the sequence of plasmids in the population (recombined or not). **8.** The experiment was repeated for each strain, and the resulting 48 vials of cryopreserved cultures were used as inoculum in the 96-well plate assay. The plate was divided into sections: top (rows A-D) and bottom (rows E-H) and left (columns 1-6), and right (columns 7-12). One stock culture was used to inoculate two wells of the plate - one in the top section and a corresponding one in the bottom section). Stocks from log exposure experiments were used in the left section, and stocks from the lag exposure experiments were used in the right section). Each quadrant is further divided vertically in half. The left part of those sections (columns 1-3 and 7-9) was inoculated with stocks pre-exposed to IPTG, and the right part of the sections (columns 4-6 and 10-12) was inoculated with the stocks never exposed to IPTG. Each pair of rows (A-E, B-F, C-G, D-H) corresponds to one strain. The plate is incubated at 37°C, shaking at 200 rpm, and fluorescence and OD600 are measured every 30 minutes. The bottom four rows are induced with IPTG (0.5mM final concentration) when OD600 in those wells reaches 0.4. The plate is then returned to the plate reader for overnight incubation and data collection.

An accurate comparison of the culture productivity between strains was possible due to applying an altered equation described in (Leveau, 2001) (full equation description is available in the methods section 2.8). In that paper, the authors include various factors affecting the final relative fluorescence culture reading in an equation to estimate the actual promoter activity, such as protein degradation rate in its expanded version. However, this can be omitted, and the simplified equation may be used when expressing stable proteins. Superfolder GFP used in this study is such a protein, resistant to proteasomal degradation (Khmelinskii *et al.* 2015). Unlike in the original paper, which considered only the exponential growth phase, the entire 24-hour growth and fluorescence data were used here. In order to do so, growth rates μ and f_{ss} were estimated individually for each half an hour of growth by fitting a straight line between the last and next ln(OD600) data point, and P was calculated for each half an hour period as well. This resulted in some noise in the data, particularly in the second half of some P(time) graphs, where P took negative values. This was caused by variations in OD600 readings over time because of the plate reader error and bubbles in the cultures interfering with the readings (which are forming during culture shaking).

In some cases, a negative slope of the growth curve for a half-hour period was calculated (which is then taken as the growth rate μ value for further calculations). Similarly, some f_{ss} values obtained for later time points took negative values as OD600 in the later time point became lower than in the previous one. These negative values were excluded from further analysis, and the baseline for the area under the curve calculations was set to 0.

Unlike other methods of estimating culture productivity over time (such as an approximation based on OD-adjusted fluorescence), this method allows accounting

for potential confounders - such as slower-growing cultures appearing brighter, resulting in higher fluorescence readings than their fast-growing counterparts, even after OD adjustments (Leveau, 2001). It also provides valuable insight into the timeline of promoter activity, which can be analysed in conjunction with growth data.

The equation produces culture productivity values, considering the protein maturation time. Therefore, the units in which the promoter activity is expressed are Relative Non-fluorescence Units (RNU) x OD600 per hour. It is essential to consider that this is an approximation method based on mathematical modelling of raw data. As such, it only provides meaningful insight when comparing datasets modelled in the same way.

3.2.6 Pre-exposure to the inducer lowers the promoter activity of pAVE011 and pIAH011-carrying strains during subsequent inductions

In order to investigate the promoter activity in all four *E.coli* strains (1515, 1516, IAH015, IAH016), the data from the experiment described fully in **Fig. 3.17** was analysed using the equation modelling approach described above. The experimental design featured inducing an array of cultures in a 96-well plate following their growth to an OD600>0.4 (re-induction). These cultures were inoculated with cryopreserved samples of the following: a) cultures inoculated into fresh media from an overnight culture, induced when their OD600>0.4 and incubated overnight (referred to as pre-exposed in log/exponential growth phase) b) cultures inoculated from a cryopreserved stock into fresh media, induced immediately and incubated overnight (referred to as pre-exposed in lag growth phase) c) control cultures grown in parallel which have not been induced (referred to as never/non-exposed controls).

Some baseline activity was observed in uninduced, never exposed to inducer strains, particularly pAVE011 carrying K background strain, 1515 **(Fig. 3.18A)**. This observation will be further investigated in the following section of this chapter. The baseline promoter activity was lower for pre-exposed strains. However, those strains also responded to re-induction with a much lower promoter activity when compared to their non-exposed controls **(Fig. 3.18)**. These baseline activity figures were later used to obtain baseline adjusted figures. The highest promoter activity was observed in strains never exposed to IPTG, which were then induced regardless of host background or plasmid compared to the same strain pre-exposed to the inducer **(Fig. 3.18)**. This describes the expected increase in the promoter activity following induction. If pre-exposure to the inducer did not affect this, the values for promoter activity of the pre-exposed, re-induced strains would be very similar to the

corresponding expected values in the same host/plasmid combination background. However, pre-exposure to the inducer caused lower promoter activity after re-induction than that of never-exposed strains. This effect was seen across genetic host and plasmid backgrounds, with the effect being more pronounced in cultures pre-exposed to IPTG in the lag growth phase **(Fig. 3.18**, **3.20** ANOVA: F (1, 8) = 84.04 P<0.0001***)*** than those pre-exposed in the exponential growth phase **(Fig. 3.19**, **3.20** ANOVA: F(1, 8)=8.254, p=0.0207***)***.

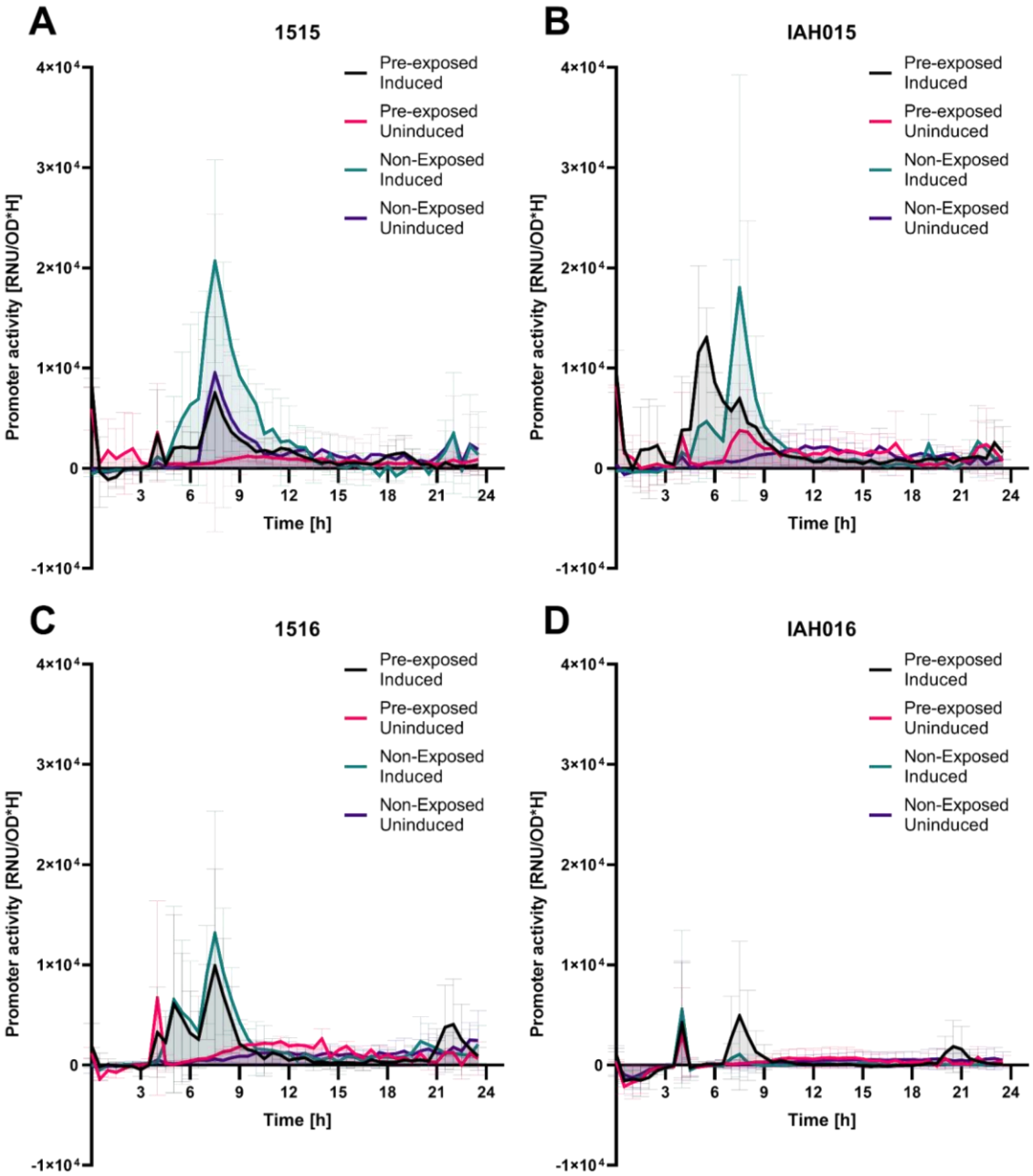


Fig. 3.18 The promoter activity of various *E. coli* strains over time (pre-exposure to the inducer in the exponential growth phase). The K background strains 1515 and IAH015 are shown in panels A and B. The B background strains 1516 and IAH016 in C and D. Pre-exposed and non-exposed refer to IPTG addition in the exponential phase of culture growth of the cryopreserved stocks from tube cultures (the 5A step of the protocol described in Fig. 3.6), whereas induced and non-induced refer to the addition of IPTG to cultures started from those stocks, inoculated in a 96 well plate. The induction time was OD600 dependent (above 0.4), usually around 3-4h. The graphs present means and SD error bars for two replicate values (means from 2 experiments calculated from triplicate values).

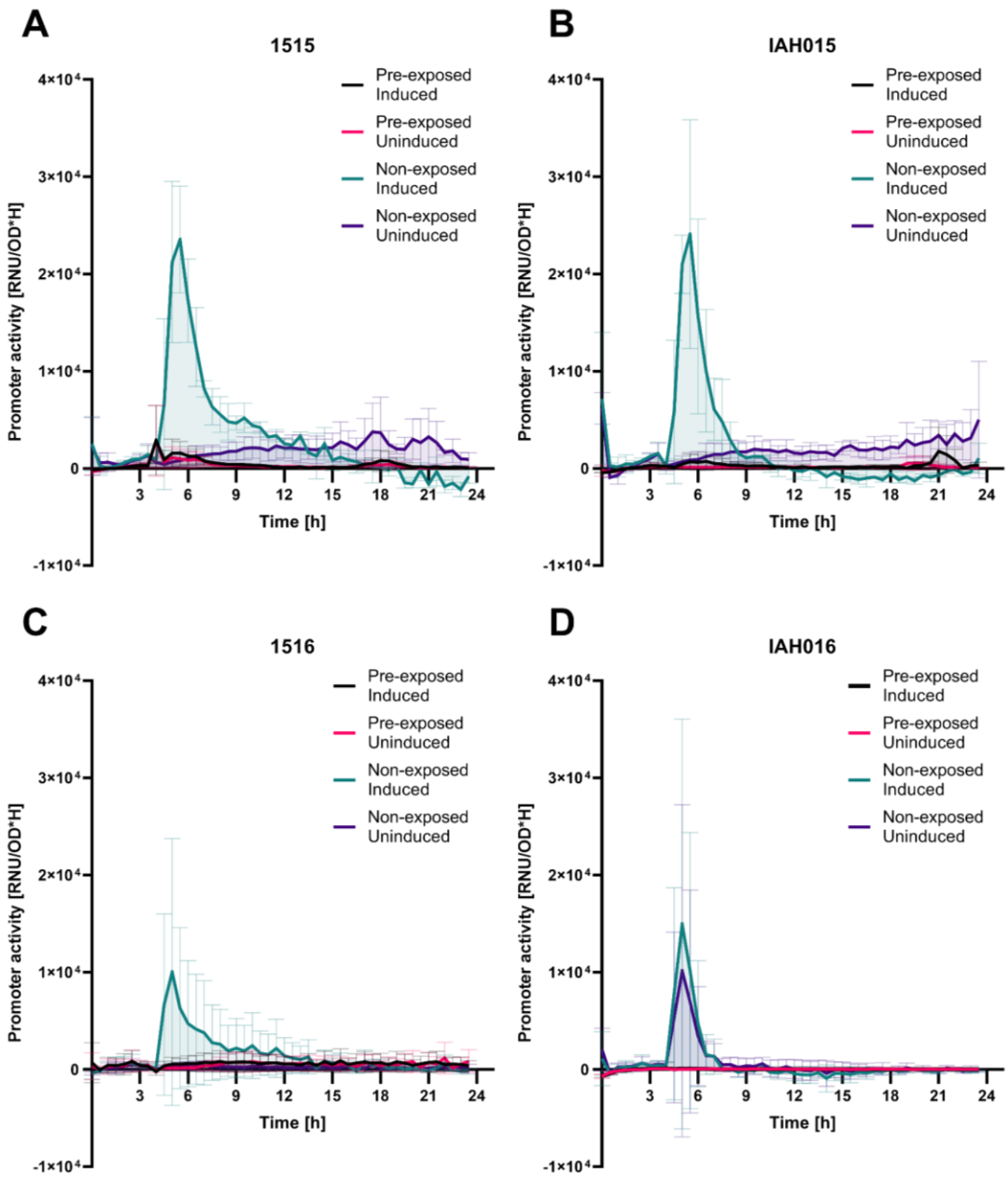


Fig. 3.19 The promoter activity of various *E. coli* strains over time (pre-exposure to the inducer in the lag growth phase). The K background strains 1515 and IAH015 are shown in panels A and B. The B background strains 1516 and IAH016 in C and D. Pre-exposed and non-exposed refer to IPTG addition in the lag phase of culture growth of the cryopreserved stocks from tube cultures (the 2B step of the protocol described in Fig. 3.17), whereas induced and non-induced refer to the addition of IPTG to cultures started from those stocks, inoculated in a 96 well plate. The induction time was OD600 dependent (above 0.4), usually around 3-4h. The graphs present means and SD error bars for two replicate values (means from 2 experiments calculated from triplicate values).

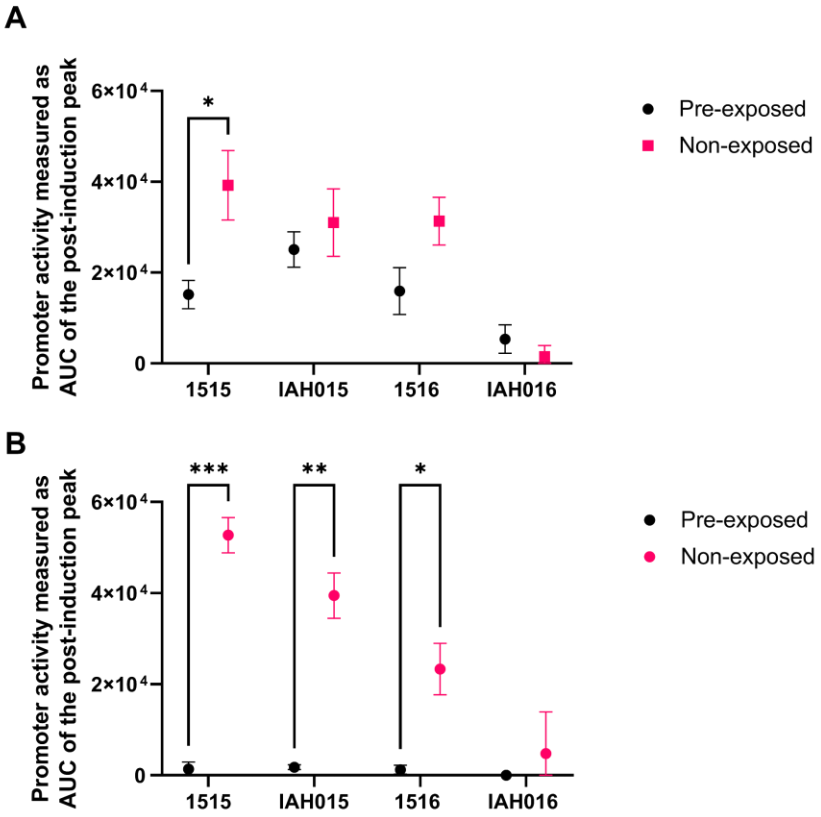


Fig. 3.20 Promoter activity change in various *E. coli* strains after induction measured as AUC of the first peak in a promoter activity over time graph. A) Cultures pre-exposed to IPTG in the exponential phase of growth protocol compared to non-exposed controls. **B)** Cultures pre-exposed to IPTG in the lag growth phase compared to non-exposed controls. Individual values represent the mean of promoter activity (for two replicate values - means of two triplicate experiments). Each mean had the relevant, experimentally established baseline activity value subtracted before plotting and analysis. The error bars show SEM. Asterisks refer to the following p-values: p≤0.05 - *; p≤0.01 - **; p≤0.001 - ***; p≤0.0001 - ****.

The lag growth phase of bacteria is a very organised, structured, and active growth phase; while the cells do not divide, transcription and translation of many genes needed to adapt to the environment and prepare for cell division take place (Bertrand 2019; Madar *et al.* 2013). These processes are crucial to the successful establishment

of the population. Therefore, it is not surprising that attempting induction in this growth phase introduces a transcriptional and translational conflict between the crucial host genes and the plasmid-introduced heterologous protein gene, especially because pAVE011 promoter T7A3 is a strong promoter. This acts in addition to the protein expression burden observable after induction in the later growth phases.

Following re-induction, the values for the promoter activity in the log pre-exposed 1515 strain were significantly lower than the values for the control, never exposed strain **(Fig. 3.20A***;* Post-hoc Sidak's multiple comparisons test, mean difference=-2.41104, 95% CI=-4.7104 to -1103, adjusted p=0.0407***)***. The same intensity of this effect was not observed for the IAH015 strain pre-exposed to IPTG in the same growth phase **(Fig. 3.20A***;* Post-hoc Sidak's multiple comparisons test, mean difference=-5.9103, 95% CI=-2.9104 to -1.7104, adjusted p=0.8990***)***. This may be due to statistical power of the experiment itself, and with more replicate data analysed, the statistical significance threshold could have been achieved. In B background strains, the promoter activity decrease of the re-induced, pre-exposed strains compared to non-induced controls was similar between the strains carrying different plasmids (pAVE011 or pIAH011) **(Fig. 3.20A** *and* **B***;* Post-hoc Sidak's multiple comparisons tests: 1516 strain - mean difference=-1.5104, 95% CI=-3.8104 to 7.6103, adjusted p=0.2375; IAH016 strain - mean difference=3.8103, 95% CI=-1.9104 to 2.7104, adjusted p=0.9756***)***.

The promoter activity differences between strains can be explained in several ways. Firstly, any changes to the promoter sequence could cause an activity change, including deletion following recombination. The promoter region structure in the stock cultures used to inoculate the 96-well plate was investigated via PCR and gel electrophoresis; none of the exponential pre-exposed cultures was recombined (apart

from one replicate of non-exposed control; data not shown, the replicate was excluded from the following analysis). This suggests that promoter deletion is not the leading cause for the apparent decrease in promoter activity in pAVE011 and pIAH011 cultures exposed to IPTG in the exponential growth phase. However, it may be more relevant for K background strains pre-exposed in the lag growth phase, as all 1515 strain replicates harboured the recombined pAVE011 plasmid (data not shown). Single base-pair changes in the promoter region could also arise and affect the polymerase binding and, subsequently, the promoter activity, as reported previously (James *et al.* 2021); however, this was not investigated by sequencing analysis. Given that substitution of the O1 modestly improved the response to re-induction of the IAH015 strain when compared to the 1515 strain following exponential phase pre-exposure **(Fig. 3.18A)**, it is possible that the recombination events took place in those cultures but were missed in PCR assays due to low abundance of the recombinant plasmids. As they can provide an advantage under induction stress conditions, following IPTG addition to the 96-well plate, the population in the well could have been overtaken by recombinant plasmid carriers during the 24-hour incubation in the plate reader.

There were differences observed between B and K background *E. coli* host response to re-induction, along with the fact that pIAH011 plasmid shows slightly improved re-induction response only in the K background strain. B and K are two distinct *E. coli* genetic backgrounds, characterised by genetic differences and divergent transcriptomics and proteomics (Yoon *et al.* 2012) and their responses to growth in high-density fermentation cultures (Marisch *et al.* 2013). Therefore, the relationship between the pAVE011 and pIAH011 plasmids and the host was expected to depend on this background.

To ensure that the pIAH011 operator modification did not have adverse effects on the strains' total biomass accumulation or maximum growth rate, the above data was also used to calculate the maximum growth rate of the strains and total biomass accumulation in a 24-hour culture. It was analysed with a three-way ANOVA. There were no significant differences between the growth rates of strains across different pAVEway operators (*lac* O1 vs *lac* O1') or pre-exposure status **(Supplementary Fig. S2)**. Another 96-well plate experiment was also performed to check the influence of IPTG concentration added at induction effect on total promoter activity. No differences were observed between strains carrying original and altered plasmid (pAVE011 versus pIAH011) in the inducer concentration needed to achieve maximum promoter activity in the 24h following induction **(Supplementary Fig. S3**, ANOVA $F(1.000, 1.000) = 3.203$, $p=0.3244$**)**. Altogether, the results show that the modification to pAVE011 plasmid operator O1, resulting in the pIAH011 plasmid, did not cause significant changes that could hinder the efficiency of the fermentation processes. The results also suggest that the pIAH011 plasmid-carrying strains may follow alternative evolutionary paths to pAVE011 in response to exposure stress. There are also differences in responses based on the host genetic background, suggesting potential divergent evolutionary pathways depending on the host genome. Therefore, it is predicted that in future evolutionary experiments, some of the adaptive mutations discovered will be strain-plasmid combination specific. While one of the mechanisms identified in pAVE011 plasmids is promoter deletion via recombination, this route is impossible in pIAH011 plasmids, which directs the compensatory mutations elsewhere. The current hypothesis is that pIAH011-carrying strains are more likely to evolve in ways that retain high protein expression than pAVE011-carrying strains. These evolutionary paths will be further explored in a later chapter of this thesis.

3.2.7 Uninduced pAVEway plasmid-carrying strains show promoter activity due to starvation response during the stationary phase

The experimental determination of culture productivity over time has revealed that even the uninduced cultures produce some sfGFP. Considering the design of the vector, it is unlikely that this is due to expression from repressor-bound DNA. Instead, it was hypothesised that this is due to lactose impurities in the media.

To test this hypothesis, strains transformed with pAVE011 sfGFP or pIAH011 sfGFP were inoculated into one of the 16 media types **(Table 3.2)**. The M9 medium was used as a control, as no lactose or other sugar impurities were expected in it. As yeast extract and tryptone are components of the LB medium, these components were also tested by adding them to the M9 medium. Vegetable tryptone was also tested, as it is used in vegetable LB recipe. Lactose addition into the media was used to show that the β-galactosidase treatment is working as intended. When added to M9 and not followed by enzyme treatment it was used to show the effect of the lactose addition into the media on the “leaky” GFP expression in the pAVEway strains. The cultures were incubated in a 96-well plate reader, and a measurement of OD600 and fluorescence was taken every 10 minutes. The productivity of cultures was calculated as described previously.

Media basic composition	Lactose addition	β-galactosidase treatment
M9 minimal media (M9)	None	Yes
		No
	0.1%	Yes
		No
M9 minimal media with yeast extract (M9Y)	None	Yes
		No
	0.1%	Yes
		No
M9 minimal media with tryptone (M9T)	None	Yes
		No
	0.1%	Yes
		No
M9 minimal media with vegetable tryptone (M9VT)	None	Yes
		No
	0.1%	Yes
		No

Table 3.2. The 16 media types used to test pAVE011 and pIAH011 plasmid-carrying strains' response to media components. All media have been supplemented with glucose to a final concentration of 0.125%, which has been determined in preliminary experiments as the highest glucose concentration not triggering catabolite repression.

The highest "leaky" expression from pAVE011sfGFP plasmid was recorded in M9 minimal media supplemented with tryptone **(Fig. 3.21).** If this was due to lactose present in the media components, β-galactosidase media treatment should significantly lower culture productivity in the absence of additional lactose supplementation. However, this was not the case. To verify that β-galactosidase treatment had worked to process any potential lactose contaminants in the media components, a control experiment was performed, where the same strains were grown in media supplemented with lactose to a final concentration of 0.1% (2.92mM). The treatment protocol was previously reported as sufficient to digest this amount of lactose in the media completely (Grossman *et al.* 1998). Here, the treatment of lactose-supplemented media reduced the culture productivity of both 1515 and 1516 strains, with the BL21 background strain still showing significant "leaky" expression in all tested media types **(Fig. 3.21)**.

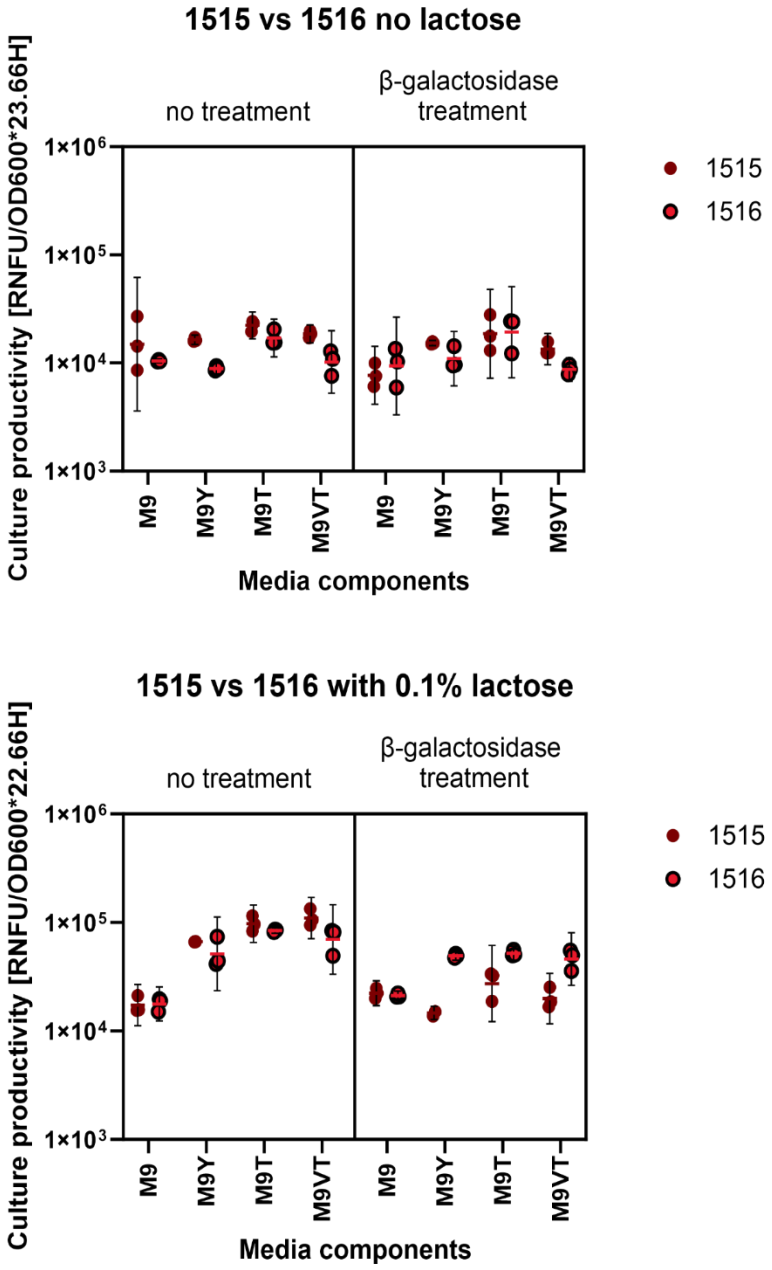


Fig. 3.21 The culture productivity of pAVE011 sfGFP strains in various media. The productivity was calculated as the AUC of Productivity over time charts for each culture. Bars represent geometric mean and 95% CI error bars. Each data point represents a culture.

To further understand the differences in responses between BL21 and W3110 (K) background strains, the experiment was repeated with pIAH011-carrying strains. Here, the highest "leaky" expression was observed in M9 media supplemented with vegetable tryptone **(Fig. 3.22)**. This directly contradicts the hypothesis of media lactose impurities affecting the expression from pAVEway plasmids, as lactose can only be present in milk-byproducts, which vegetable tryptone does not contain. Furthermore, similarly to the responses shown by pAVE011-carrying strains, there is a significant protein expression in M9 minimal media. Therefore, the "leaky" expression observed must be caused by other factors.

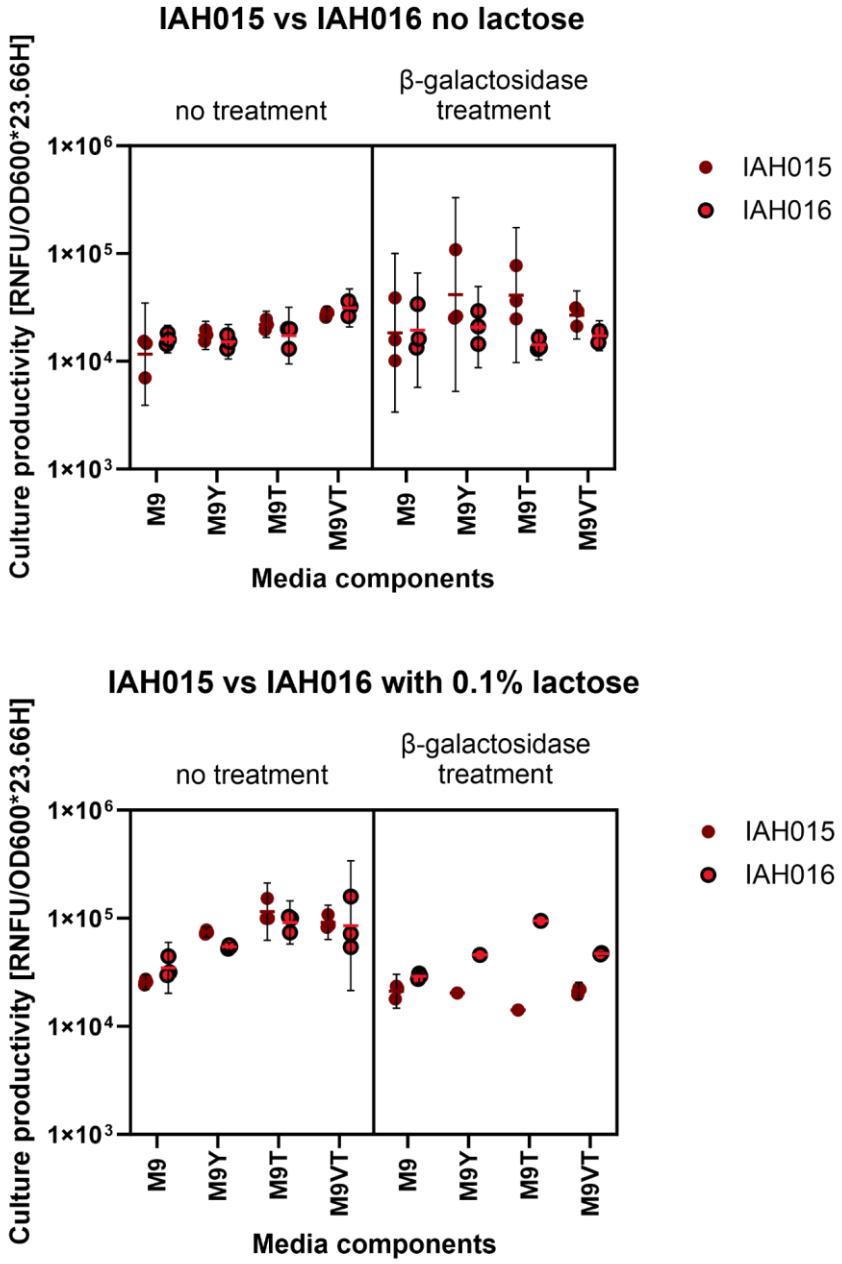


Fig. 3.22 The culture productivity of pIAH011 sfGFP strains in various media. The productivity was calculated as the AUC of Productivity over time charts for each culture. Bars represent geometric mean and 95% CI error bars. Each data point represents a culture.

Both tryptone and vegetable tryptone are used as sources of nitrogen. Meanwhile, yeast extract is a more complex component, providing carbohydrates, amino acids and peptides. Similar results have been previously reported (Grossman *et al.* 1998), where high recombinant protein overexpression in rich YT media was observed. The authors measured the lactose content of the media to be below 0.002% and pre-treated the media with β-galactosidase prior to inoculation, which had no impact on the protein overexpression. Thus, they eliminated the possibility that the lactose impurities present in the media were responsible for the heterologous protein overexpression. Instead, they demonstrated that the phenomenon was linked to the stationary growth phase of the bacteria (the starvation response) and that acetate and cAMP were required to reproduce this effect. Meanwhile, glucose supplementation has been shown to reduce the effect.

The BL21 background strains, regardless of the pAVEway plasmid variant carried, have shown higher culture productivity in media supplemented with lactose and subsequently treated with β-galactosidase compared to the K background strains **(Fig. 3.21, 3.22).** As highlighted in the earlier sections of this chapter, *E. coli* strains B and K differ not only genetically but also in their responses to environmental factors. B background *E.coli* is characterised by higher sensitivity to specific environmental stressors (Yoon *et al.* 2012), which may account for the observed differences in “leaky” protein expression from pAVEway plasmids in nutrient-deficient media types.

In order to investigate the effect of glucose supplementation on culture productivity, pAVEway plasmid-carrying strains were grown in vLB, which was supplemented with glucose to final concentrations of 0%, 1% and 0.25%. In a preliminary experiment, the 0.25% glucose concentration was determined as the lowest glucose concentration capable of reducing leaky expression in the absence of

an inducer **(Supplementary Figure S4)**. This concentration, 1% and a control 0% (non-supplemented vLB) were then chosen to test the uninduced and induced expression in pAVEway plasmid-carrying strains.

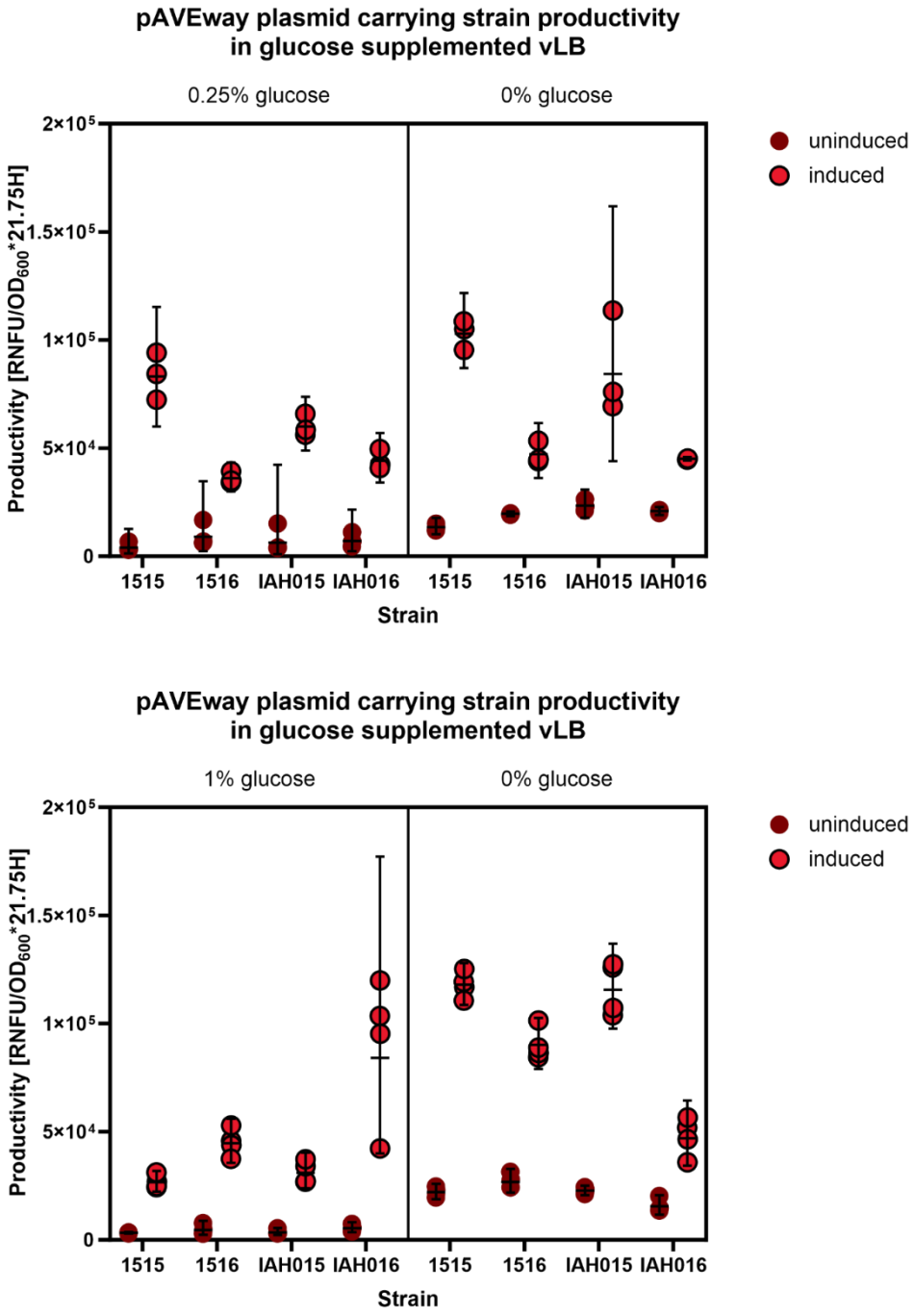


Fig. 3.23 The cumulative productivity of pAVEway plasmid-carrying strains calculated over 21.75h of induced growth. The top panel shows differences in productivity between strains grown in media with no additional glucose supplementation and strains grown in media supplemented with glucose to a final concentration of 0.25%. The bottom shows data from a repeat experiment, where media was supplemented with glucose to a final concentration of 1%. Each data point represents a replicate; geometric means and their 95% CI are shown with the bars.

The addition of glucose to the culture media results in a lower uninduced response in pAVEway plasmid-carrying strains **(Fig. 3.23)**. While 1% glucose supplementation results in suppressing the uninduced heterologous protein production up to 6.7 times when compared to non-supplemented media, it also results in up to 4.4 times lower total yield upon induction. In contrast, glucose supplementation to final 0.25% in the growth media results in up to 3.8 times reduction in uninduced heterologous protein expression compared to non-supplemented media while having minimal effect on the induced yield **(Table 3.3)**.

Glucose supplementation	Strain	Uninduced productivity reduction (geometric mean with 95% CI)	Induced productivity reduction (geometric mean with 95% CI)
0.25%	1515	3.381 times (0.83 to 13.771)	1.237 times (0.0787 to 1.946)
	1516	2.189 times (0.583 to 8.229)	1.311 times (1.207 to 1.423)
	IAH015	3.768 times (0.528 to 26 898)	1.405 times (0.898 to 2.196)
	IAH016	2.958 times (0.899 to 9.727)	1.022 times (0.795 to 1.314)
1%	1515	6.709 times (5.67 to 7.939)	4.423 times (3.794 to 5.156)
	1516	5.923 times (3.504 to 10.011)	2.018 times (1.605 to 2.537)
	IAH015	6.521 times (4.235 to 10.040)	3.728 times (3.364 to 4.132)
	IAH016	2.911 times (2.181 to 3.885)	0.559 times (0.266 to 1.175)

Table 3.3. Calculated geometric means of productivity reduction of pAVEway plasmid-carrying strains. The productivity reduction was calculated by dividing the productivity in un-supplemented media by the matching replicate productivity in supplemented media. The calculations were done in GraphPad Prism 10.0.2.

This result could be explained by lactose in the vLB media and catabolite repression due to glucose addition. However, the previous experiment has shown this to be an unlikely cause. The alternative hypothesis is that the leaky expression is a response caused by starvation coinciding with the culture entering the stationary growth phase. The responses of uninduced cultures over time **(Supplementary Figure S5)** show that the leaky expression response usually starts in the mid to late log phase or at the beginning of the stationary phase. The same trend has been observed in previous experiments **(Fig. 3.17, 3.18)**.

These results are consistent with previously reported data (Grossman *et al.* 1998) on spontaneous protein expression dependent on carbon metabolism. However, this is not the only starvation response linked to heterologous protein production. More recently, *E. coli* starvation response (specifically phosphate starvation) has been linked to lactose metabolism and expression of *lac-controlled* heterologous proteins (Pandi *et al.* 2020), and this discovery has been used in the design of novel auto-inducible protein expression systems (Gundinger *et al.* 2022; Menacho-Melgar *et al.* 2020).

In this chapter, the pAVE011 vector was characterised in two genetic host backgrounds. Its intrinsic instability was investigated and found to be a result of RecA-independent recombination between the two homologous operators surrounding the promoter. The recombination can be induced in pAVE011 plasmid-carrying strains by exposure to the inducer in early growth phases of the culture, but not in pIAH011 plasmid-carrying strains. Following the recombination, the heterologous gene expression is not completely lost, but markedly lowered. However, homologous recombination between the plasmid's operators is not the only mechanism responsible

for the productivity drop after exposure to inducer. The uninduced “leaky” expression from the pAVE011 and pIAH011 plasmids was also investigated and was found to likely be caused by the starvation response triggering in the later stages of the culture growth. Finally, a method for culture productivity calculation and comparison is introduced.

The characterisation of pAVE011 and pIAH011 plasmids presented in this chapter provide a valuable foundation for the evolutionary work presented in the next chapters.

4. Results II “Development and application of a novel FACS-based selection of *E. coli* strains with improved recombinant protein production capabilities

4.1 Introduction

Despite limitations discussed in introduction, *Escherichia coli* remains one of the industry workhorses for recombinant protein production. It is an attractive host due to its ease of genetic manipulation, rapid growth and cheap growth media. However, some of *E. coli*'s characteristics make it less suitable for certain fermentation processes, which have recently gained popularity in the protein production industry, namely repetitive fed-batch and continuous fermentation.

The most commonly used protocol with *E. coli*, fed-batch manufacturing, consists of two distinct phases: biomass accumulation and induction. During the first part of the process, expression of the recombinant protein is inhibited while the host culture reaches desired biomass. Expression of the recombinant gene is then induced, and the cells produce the target protein over a defined number of hours, after which the product is harvested. Repetitive fed-batch and continuous fermentation protocols aim to increase the overall yield of the manufacturing process by prolonging the time during which the host is induced and producing the target protein (Liu et al. 2020). Continuous fermentation is the most attractive from the efficiency perspective, as it only requires a single batch set-up, from which the product is continuously harvested. At the same time, fresh nutrients are provided for the growth of the remaining culture (Li et al. 2014). In contrast, during repetitive batch fermentation, the product is harvested periodically while the remaining culture is provided with fresh feed and immediately induced (Kopp et al. 2020). These approaches improve yields and save

the time it would take to prepare the equipment between several separate fed-batch fermentation runs.

While continuous fermentation protocols are not yet widely used, they are attracting interest from the industry as an alternative to traditional fed-batch protocols (Li et al. 2014; Chen and Jiang 2018), the characteristics of pAVEway plasmids make them unsuitable for these approaches. Firstly, pAVE011 plasmids undergo homologous recombination of the promoter region at high rates when hosts are exposed to the inducer too early in their growth, substantially reducing the yield (see sections 3.2.1 and 3.2.6). Secondly, when no recombination occurs, both pAVE011 and pIAH011 plasmids show reduced yields upon secondary induction, even if appropriate time is allowed before induction for cells to reach the exponential growth phase. These issues are likely related to the burdens to the cell of both recombinant protein production and plasmid maintenance, as pAVEway vectors use a very strong IPTG inducible T7A3 promoter.

The relationship between the plasmid and its host must be improved to reduce the burden of plasmid carriage and protein overproduction without sacrificing overall yield. It has been previously reported that the plasmid carriage burden varies based on specific host and plasmid combinations and that the relationship between the plasmid and its host can be improved through evolutionary experiments. The coevolution between plasmid and bacteria can be directed towards a more stable phenotype by using predefined selection pressures.

There are many approaches to select for beneficial characteristics, and the one employed in this study is fluorescence-activated cell sorting (FACS). FACS recently gained popularity as a tool for experimental evolution due to its highly precise mechanism of selecting single cells based on their phenotype (Tracy et al. 2010).

Several applications have been investigated: evolving larger bacteria (Yoshida et al. 2014; Tian et al. 2023), evolving enzymes with enhanced activity produced by bacteria (Yang & Withers 2009) and, changing the bacterial affinity for specific substrates (Olsen et al. 2003). Most relevant to this project, it has been previously used to successfully enhance plasmid stability in a bacterial population (Deatherage et al. 2018).

The work in this chapter describes the design and implementation of FACS-based directed evolutionary approach selecting for plasmid carrying *E. coli* clones showing high GFP activity while retaining tight transcriptional control over induction of expression. The obtained evolved *E. coli* lines were then assessed in terms of productivity and protein expression stability and compared to their ancestors. The results suggest that positive selection for high protein production can be used as a proxy to select for *E. coli* strains with improved biotechnological characteristics, even though the evolutionary outcomes can be variable depending on the *E. coli* genotype, plasmid identity and replicate selection line.

4.2 Results and discussion

4.2.1 Parameterising a FACS-based selection screen

The evolutionary experiment was designed to include several environmental pressures. An overview of the workflow detailed in this section is shown in **Fig. 4.1.**

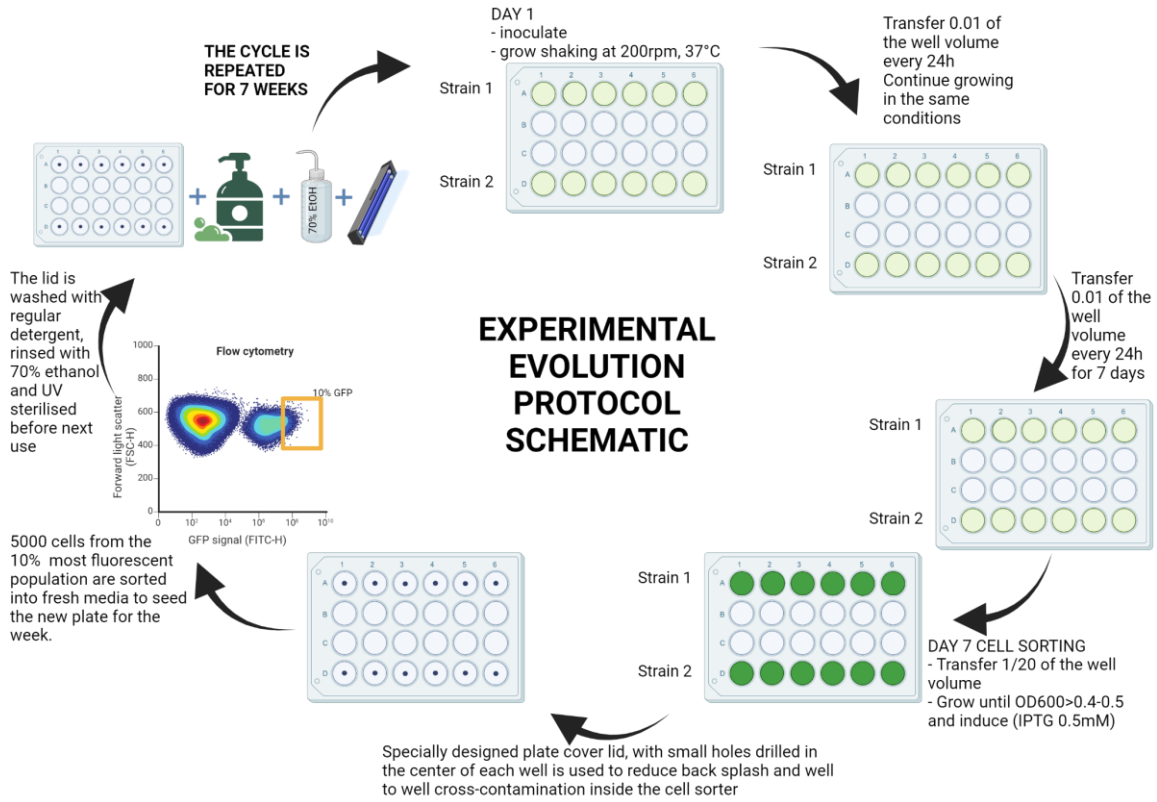


Fig. 4.1 Experimental evolution protocol schematic. Four plates were set up on day one, each containing a pair of strains inoculated in 6 replicates in rows A and D. The pairs were: 1515 and IAH015, 1516 and IAH016, empty vector carrying BL21 and W3110 *ΔompT,* and no vector carrying BL21 and W3110 *ΔompT.* For the next seven days, a small volume (1/100) of each culture was transferred into fresh media (supplemented with tetracycline to a final concentration of 10 µg/ml where relevant). The plates were continuously shaking at 200 rpm in 37℃. Once every week, the transfer volume was increased to facilitate faster growth to an appropriate OD600 for induction with IPTG (final concentration 0.5 mM). After several hours of incubation following induction, the cells were sorted into fresh media containing plates using a fluorescent cell sorter. The cells selected were from the top 10% most fluorescent population. These plates were then used as the starting plate in the serial transfers

performed next week. The cycle was repeated a total of 7 times, or approximately 328 generations. Schematic created with biorender.com.

First, the growth media was chosen to simulate as closely as possible the growth media the production strains are most likely to be grown during the industrial processes to ensure that any media-specific growth adaptation will not affect the strain's performance. The selection antibiotic for the pAVEway vector, tetracycline, was also included to ensure no plasmid loss and encourage alternative evolutionary pathways towards higher induction and recombinant protein production stress tolerance. The experimental design included two types of bottleneck events: the non-specific daily selection and the high fluorescent signal-dependent weekly selection. The daily selection favoured enhanced growth, as random transfer of a small volume of the culture (1/100) was more likely to contain highly abundant, well-growing cells. The weekly cell sorting events specifically selected the cells with the highest fluorescent signal, which was chosen as an approximation of productivity.

One of the main concerns identified during preliminary FACS experiments was the risk of cross-contamination. This was discovered when bacterial growth was observed in negative control wells filled with sterile media on the same plate into which bacterial cells were sorted using the FACS machine. Therefore, the experimental design included cross-contamination prevention steps in both the plates design and handling and cleaning the FACS machine during sorting.

Firstly, four 24-well plates were set up on the first day of the evolutionary experiment. Two strains were inoculated in each plate (rows A and D) in 6 biological replicates, leaving wells in rows B and C empty and acting as separators between the strains on the same plate. The stocks used to inoculate the plates on the first day were not clonal, but mixed cryopreserved populations. Secondly, each plate contained a

pair of strains chosen based on the hypothetical impact of potential cross-contamination on the final experiment results.As it was more important to prevent contamination between the two genetic *E. coli* backgrounds carrying the pAVEway plasmids than between plasmid variant carried, strains B and K were separated onto different plates. This is because in the event of cross-contaminating cultures of one genetic background with the other, the risk of total loss of one of them from the population due to competition is high. Similarly, to prevent contaminating plasmid-free cultures with empty plasmid-carrying cells, the two types of controls were separated onto two plates, as contaminating empty-plasmid carrying cultures with plasmid-free cells would likely lead to plasmid loss due to competition. Therefore, W3110 *ΔompT* strains carrying pAVE011 or pIAH011 plasmid (1515 and IAH015) were paired on the first plate. BL21 strains carrying either pAVE011 or pIAH011 plasmid (1516 and IAH016) were paired on the second plate. The third plate contained no vector control strains BL21 and W3110 *ΔompT,* while the last plate contained an empty vector (pAVE011) carrying control strains BL21 and W3110 *ΔompT.*

The media used was vegetable LB (vLB) with the tryptone component substituted with vegetable tryptone in the recipe. All cultures except the no vector control media also contained tetracycline at the final 10 µg/ml concentration. The culture volume was 2 ml, and the plates were continuously shaken at 200 rpm at 37°C. Once every day (except for fluorescent cell sorting selection days), the plates were taken off the shaker, and 20 µl of each culture was used to inoculate fresh media in 24-well plates.

The transfer protocol was altered once a week. 100µl of each culture was transferred into fresh media in 24-well plates and grown at 37°C, shaking at 200 rpm for 1.5-2 h. When the cultures reached OD600 of 0.3-0.5, IPTG was added to a final

concentration of 0.5 mM to all four plates (including controls). The cultures were left on the shaker to grow and express the fluorescent protein, usually for three hours. Immediately before flow cytometry, the samples were diluted with fresh vLB media to ensure low enough concentration needed for fluorescent cell sorting.

The bacterial populations in samples were identified by measuring their forward and side scatter (FSC and SSC) and compared to measurements taken of sterile vLB media **(Fig. 4.2A)**. Forward scatter indicates the relative size of the cell, while side scatter indicates the granularity of the cell. In sterile vegetable LB, multiple events are recorded **(Fig. 4.2A)** due to media complexity and its components. On a scatter plot depicting FSC and SSC of negative control cells in vLB (W3110 *ΔompT,* **Fig. 4.2B**), multiple events similar to those observed in sterile vLB are recorded, as well as an additional dense cluster, identified as the bacterial population. The same cluster is also observed in a sample containing strain 1515, (W3110 *ΔompT* pAVE011) **(Fig. 4.2C).** These results allowed for the creation of the first gate, isolating bacterial population events from events caused by media components.

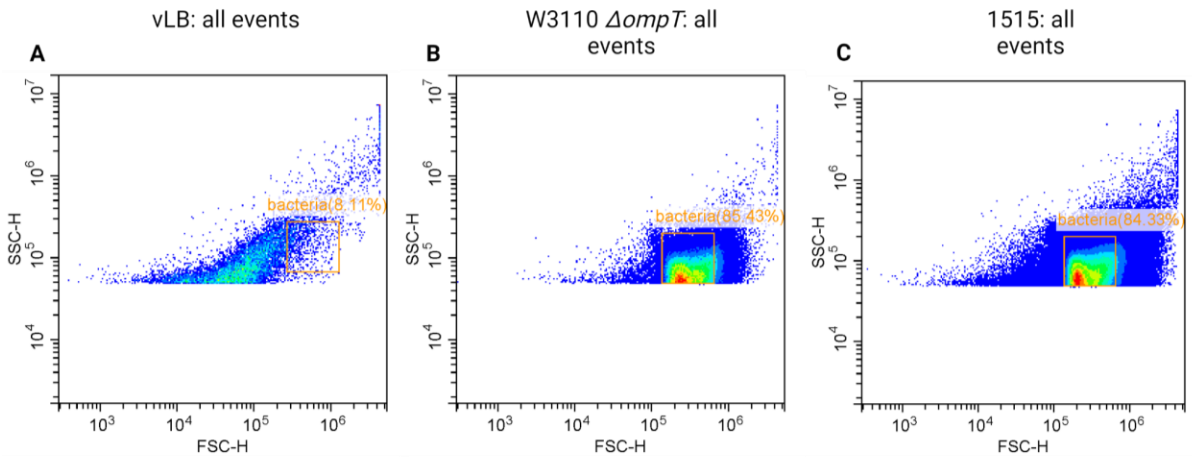


Fig. 4.2. Exemplary data of the identification process of bacterial populations in complex media using their forward and side scatter qualities. A) FSC and SSC of sterile vLB media **B)** FSC and SSC of W3110 *ΔompT* negative control in vLB **C)** FSC and SSC of strain 1515 in VLB. The orange box in panel **C)** depicts the bacterial population gate established based on clustering data points in panel **B).** The orange box in panel **A)** was established based on clustering of data points in another control sample (data not shown)

Within the identified bacterial population, the green fluorescent signal was measured and plotted against FSC **(Fig. 4.3A, B)**. The fluorescence of W3110 *ΔompT* negative control was 10^3 **(Fig. 4.3A)** compared to that of the pAVE011-*gfp* plasmid-carrying strain 1515 – 10^5 **(Fig. 4.3B).** These two populations are clearly separated. Using a histogram of the fluorescent population of the plasmid-carrying bacteria, the top 10% of the fluorescent bacterial population was marked for selection from each of the six cultures of each of the four experimental lines that expressed GFP (1515, 1516, IAH015, IAH016; example shown in **Fig. 4.3C)**. Five thousand events (bacterial cells) were sorted from this high expression sub-population into new media on a 24-well plate. During sorting, the data on population heterogeneity was also collected.

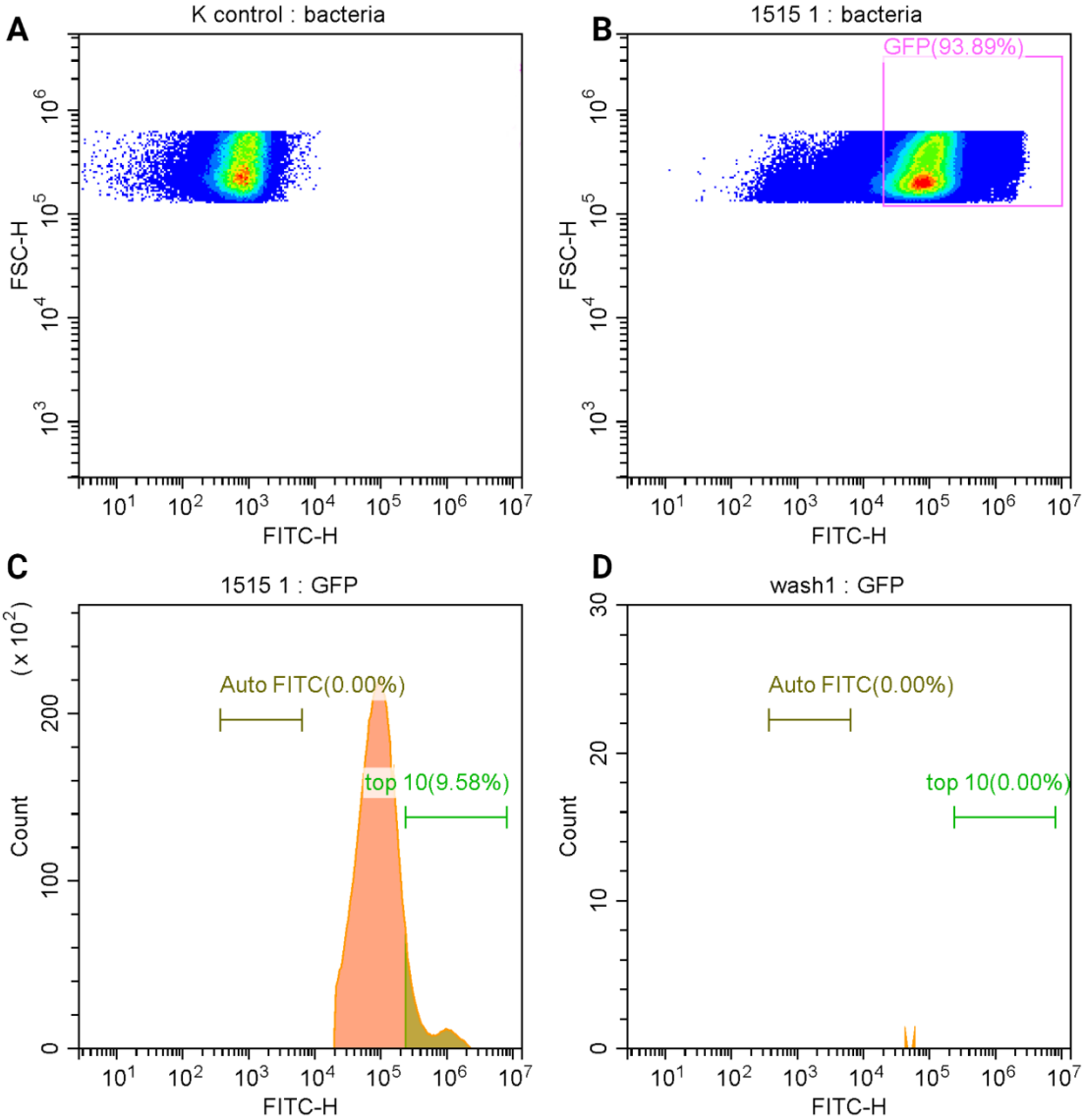


Fig. 4.3. An exemplary dataset illustrating the 10% sub-population sorting. A) The fluorescence (FITC-H, x axis) of K background control strain (W3110 *ΔompT*) plotted against the size of the cells (FSC). **B)** The fluorescence of W3110 *ΔompT* pAVE011 sfGFP strain (1515) plotted against the size of the cells (FSC). **C)** The histogram of green fluorescent cells within the 1515 strain population. **D)** The histogram of water used in the last step of the wash cycle between the culture samples.

To reduce cross-contamination during FACS, several steps were added to the cell sorting protocol. Firstly, the plate containing the fresh media that bacteria were sorted into was at all times covered with a modified plate lid of my own design, which had precisely drilled holes in the centre of each well. This allowed the stream of selected cells to fall into the well but prevented the contents of the well from splashing back up and transferring into neighbouring wells. It also reduced the risk of environmental contamination. This modified lid was wiped with 70% ethanol after sorting each strain. It was reused each week, and each time before using, it was washed with water and detergent, wiped with 70% ethanol, and placed in a laminar flow hood cabinet under UV light for 1 hour, together with an open empty 24-well plate packaging. It was then placed in this packaging (inside the cabinet) and closed using UV sterilised tape.

A washing protocol of the fluorescent cell sorter lines was also followed after each sample was sorted to ensure no contamination between cultures run through the cell sorter one after another. Filter sterilised detergent was placed in 4 tubes, as was filter sterilised water and bleach. Each tube was assigned to one of each strain. After each bacterial sample was run and sorted, the detergent was run for several minutes, followed by the water, which was also run for several minutes. When switching between strains, the bleach solution was run through the lines of the cell sorter for at least 2 minutes before detergent. All wash steps were performed using high sample flow speed to ensure timely washing of the instrument lines. Due to time and technical constraints, removing all bacteria during the washing step was not always possible. Therefore, the wash cycle was carried out until no GFP signal was detected in the previously identified top 10% subpopulation **(Fig. 4.3D).**

4.2.2 Monitoring population-level changes during the evolutionary experiment

The results of the weekly sorts were analysed, focusing on three different measurements **(Fig. 4.4).** The fluorescent fraction of the population allows for an approximation of the plasmid stability changes within the population after every week of evolution. The measurements of median fluorescent values of the population allow for monitoring the changes in the intensity of the signal (the strain productivity) over the weeks. Finally, measuring the median fluorescence of the top 10% highest producers provides information about the intensity of the fluorescent signal in cells selected to grow for the next week.

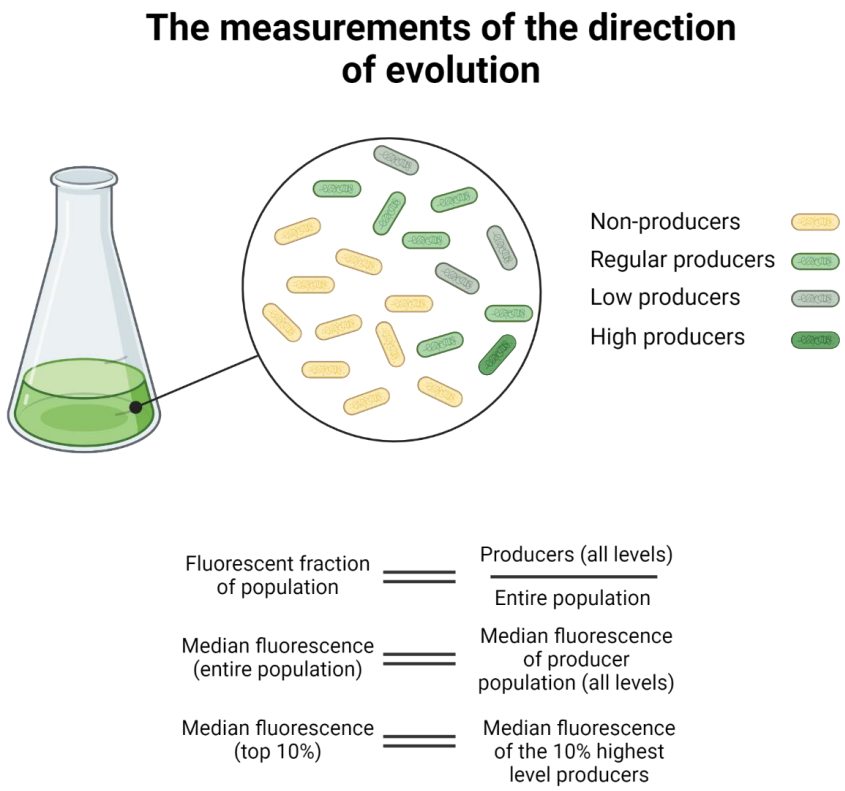


Fig. 4.4 A schematic representation of a hypothetical evolved culture population structure. Each population contains both non-producers - cells that have lost the ability to express GFP entirely - and producers. Producers exhibit a range of intensity of the fluorescent signal from very low to high.

To calculate the fluorescent fraction of the population, the number of producers is divided by the number of all bacteria in the population. The median fluorescence values were also measured for both the entire fluorescent population and the top 10% of fluorescent bacteria within the fluorescent population.

4.2.2.1 Population fluorescence fluctuates during the evolutionary experiment

The evolutionary experiment was carried out as outlined in an earlier section and **Fig. 4.1**. Six populations of four strains (IAH015, IAH016, 1515 and 1516) were followed for 7 weeks. The plasmid stability in these cultures was approximated using the fluorescent fraction of population.

All six cultures of the strain harbouring the original pAVE011 vector in W3110 genetic background (1515) did not achieve increased plasmid stability within the population after seven weeks of the experiment **(Fig. 4.5A)**. However, it is important to note that the initial fluorescent fraction of the population for this strain was around 90% or more in most cultures. In contrast, all 1516 (BL21 background) strain cultures have shown an increased tolerance of the plasmid within the population as measured by the fluorescent fraction of the population (ranging from a 0.03 increase in culture 5 to a 0.24 increase in culture 4, **Fig. 4.5B**). This improvement may be partly attributed to the lower plasmid stability starting point of these populations, with an average of 80%.

In pIAH011-carrying strains, different trends were observed. Several cultures of the IAH015 strain - 1, 2 and 3 **(Fig. 4.5C)** - have a higher fluorescent fraction of population on week 7 when compared to week 1, with cultures 1, 2 and 3 showing a fraction increase of 0.07, 0.11 and 0.3, respectively. In culture 4 **(Fig. 4.5C, 4.6C, 4.7C)**, the fluorescent fraction of the population rose throughout the experiment, with a sharp drop in week 7. An adaptation likely occurred between weeks 6 and 7, which has caused the non-producer cells to grow faster leading to producer cells becoming lost during daily transfers into new media. In the IAH016 strain, BL21 carrying pIAH011 plasmid, the plasmid stability at the beginning of the experiment was low, below 15%

for all populations **(Fig. 4.5D)** Throughout the experiment, the plasmid stability rose, reaching almost 50% in two populations - 4 and 5.

The most significant advantage of the FACS method applied is illustrated by the fluctuation of the fluorescent fraction of population, especially in the IAH015 strain **(Fig. 4.5C).** The plasmid prevalence in the population dropped significantly between weeks 2 and 3 in all populations. Without the application of the FACS-assisted selection method, the likely outcome of this trend would be complete plasmid loss. However, following the selection in week 3, the fraction of the fluorescent population rose in all cultures and in the following week 4, it reached the original levels. Additionally, selecting for fluorescence ensures selection of clones producing the target protein at high levels. This is another advantage of this approach, as alternative approaches, such as antibiotic selection, only ensure the resistance marker stability. During a long-term evolutionary experiment, the plasmid may be altered to reduce the burden to the host cell by deletion of the heterologous gene coding sequence or its alteration resulting in a truncated product. Furthermore, the antibiotic resistance gene my become integrated into the genome of the host, which would result in plasmid loss from the population.

Importantly, the “plasmid stability” discussed above refers only to the plasmid maintenance within the population. It is impossible to make conclusions about plasmid stability on a single cell level. Both stabilised and unstable plasmid variants may be present in any population at a given time (Hülter et al. 2020; Hughes et al. 2012). The fluorescent fraction of population only provides information about the percentage of population which at a given time carries a fully functional plasmid, from which the sfGFP is expressed.

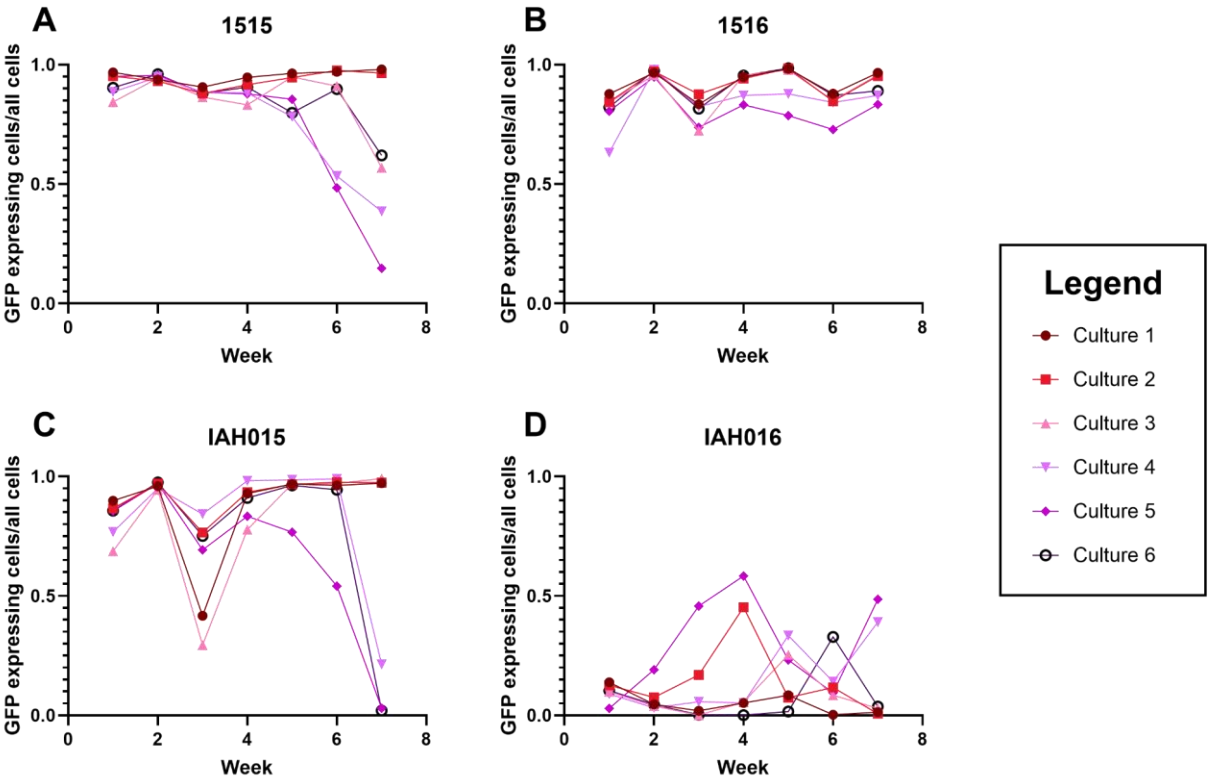


Fig. 4.5 Plasmid maintenance across 24 *E. coli* populations throughout the 7 week evolutionary experiment. Plasmid maintenance was estimated as a fraction of total population expressing sfGFP. **A)** Fraction of sfGFP expressing cells in the population of W3110 *ΔompT* pAVE011 cultures **B)** Fraction of sfGFP expressing cells in the population of BL21 pAVE011 cultures **C)** Fraction of sfGFP expressing cells in the population of W3110 *ΔompT* pIAH011 cultures **D)** Fraction of sfGFP expressing cells in the population of BL21 pAVE011 cultures. Measurements for each culture were taken during weekly FACS.

4.2.2.2 Fluorescence intensity of all four strains drops in the first week of evolution

During the evolutionary experiment, the data on population median fluorescence was also collected as described in section 4.2.2. Fluorescence intensity of a population provides information about the productivity of the plasmids on a population level, which does not correlate linearly with the percentage of the population still carrying the plasmid. The drop or increase in fluorescence may be related to plasmid carriage, but also to mutations on the plasmid or the genome, such as changes to sfGFP coding sequence.

In both strains 1515 and IAH015, there is a marked decrease in median fluorescent values between weeks 1 and 2 **(Fig. 4.6A, C)**. This was previously observed as a decreased culture productivity following repeated induction (see section 3.2.6) . Similarly to earlier observations, strain 1515 lost more of its original fluorescence signal when compared with strain IAH015 between weeks 1 and 2. The B background strains (1516 and IAH016) have also shown a marked drop in fluorescent signal between weeks 1 and 2 **(Fig. 4.6B, D).** However, by the end of the experiment, most of the cultures of strain 1516 are showing an upwards trend, suggesting a recovery in productivity. Meanwhile IAH016 strain cultures show productivity at week 7 similar to their original week 1 values, with culture 1 achieving higher total fluorescence by the end of the experiment when compared with the starting point.

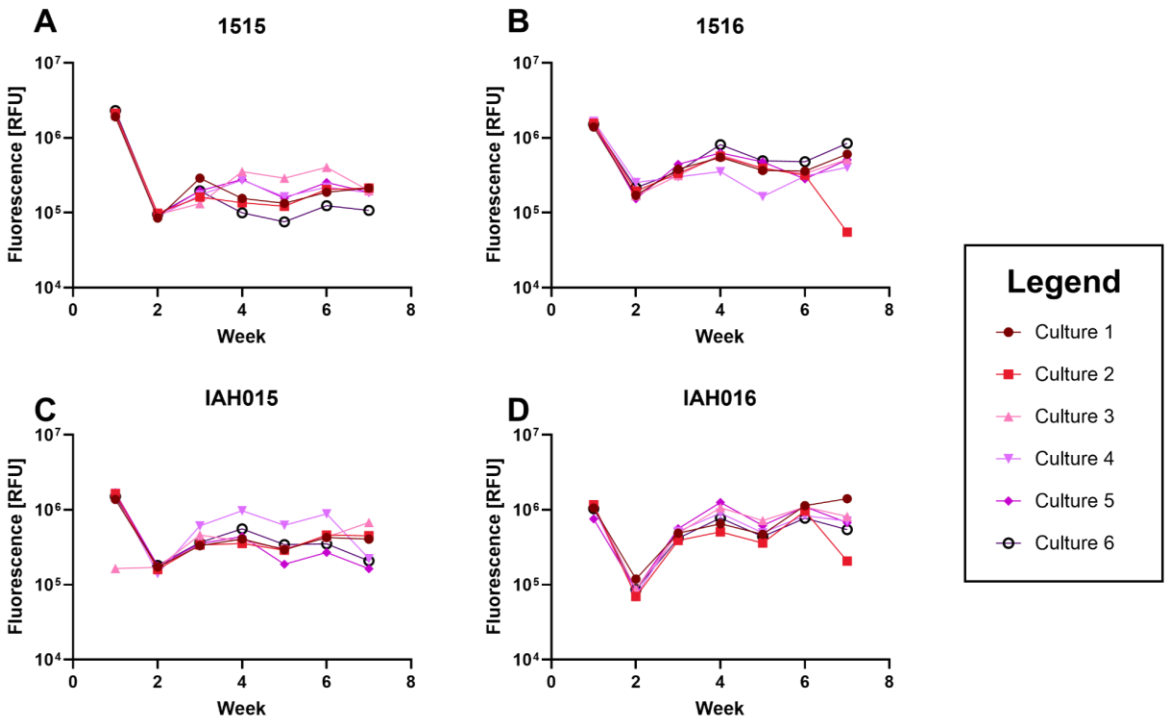


Fig. 4.6 Total median fluorescence across 24 *E. coli* populations throughout the 7 week evolutionary experiment. A) Fluorescence of sfGFP expressing cells in the population of W3110 *ΔompT* pAVE011 cultures **B)** Fluorescence of sfGFP expressing cells in the population of BL21 pAVE011 cultures **C)** Fluorescence of sfGFP expressing cells in the population of W3110 *ΔompT* pIAH011 cultures **D)** Fluorescence of sfGFP expressing cells in the population of BL21 pAVE011 cultures. Measurements for each culture were taken during weekly FACS.

A variety of factors can explain the above results. Firstly, it is possible that upon exposure to the inducer, the plasmid's promoter region is deleted via homologous recombination between cell sorting events, resulting in reduced productivity. However, this is unlikely in this case, as it was previously shown that the promoter region recombination events following exposure to the inducer in the exponential growth phase are rare. Furthermore, it was also shown that the altered operator considerably reduces this rate of promoter loss. However, here we observe strain IAH015 also losing the fluorescent signal strength between weeks one and two (**Fig. 4.6C**). As the

selection antibiotic (tetracycline) was present throughout the experiment and the population fraction of producers was high in week 2 for both strains, complete plasmid loss is also not a probable cause of the loss in fluorescence **(Fig. 4.5).** Instead, one of the following reasons may be the cause for this observed change.

Mutations in the coding sequence of recombinant proteins have been previously identified as reducing the metabolic burden of their production (Rahmen et al. 2015). However, the exact mechanism was not identified - codon usage was among the hypothesised factors. In the case of green fluorescent protein, mutations can lead to altered excitation and emission spectra, increase or decrease in brightness or folding efficiency (Pédelacq et al. 2006; Heim & Tsien 1996). Potential mutations in the sfGFP gene alleviating the protein production burden could simultaneously create low-brightness variants and explain the decrease in fluorescence of the cultures between weeks one and two. This hypothesis is also supported by same trends observed in the top 10% fluorescent population changes **(Fig. 4.7)**

Alternatively, it is also possible that the mutations arose in the origin of the replication (*ori*) region of the plasmid, reducing the protein production burden via plasmid copy number decrease. Point mutations in the promoter region reducing the affinity to polymerase are also possible. Both of these adaptations were previously reported in a GFP overexpressing BL21 (DE3) bacteria (James et al. 2021) as a response to induction stress and could also account for the fluorescence change in pAVEway vector-carrying strains between weeks one and two. These proposed mechanisms can be crucial in the coevolutionary process required for any subsequent adaptive mutations; an initial reduction in expression burden might be required for other adaptations to emerge.

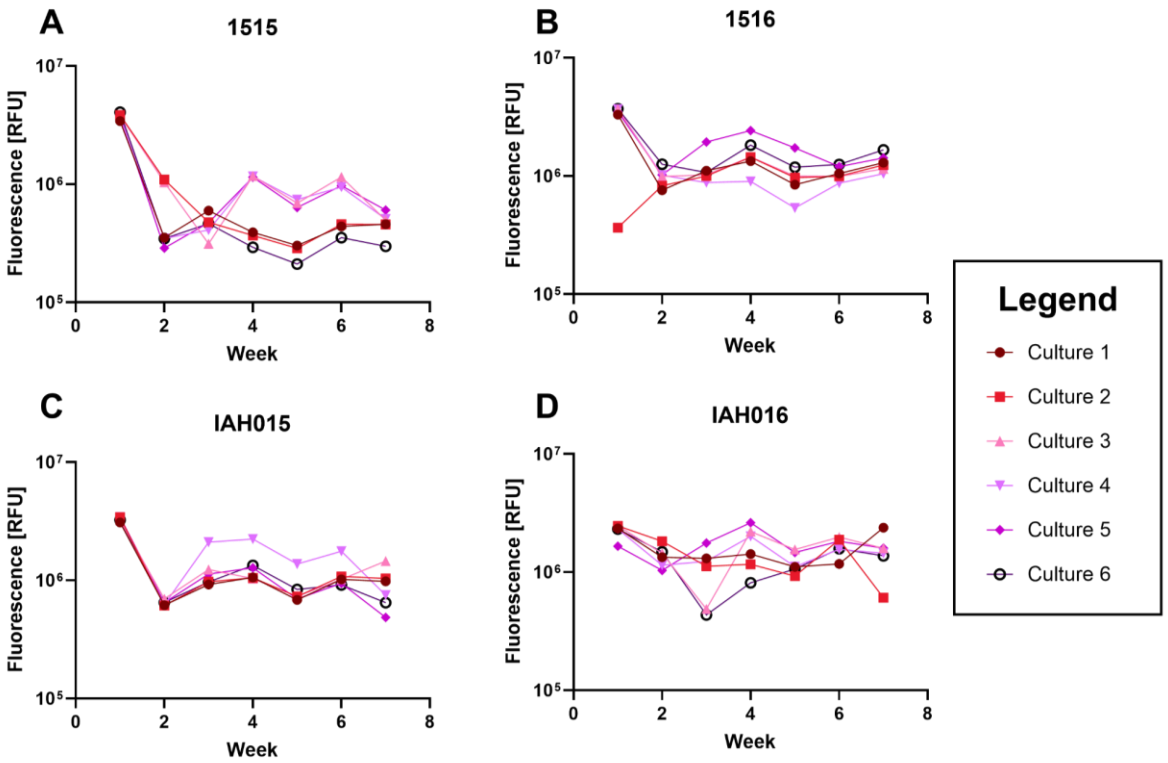


Fig. 4.7 Top 10% producer fluorescence across 24 *E. coli* populations throughout the 7 week evolutionary experiment. A) Fluorescence of top 10% sfGFP expressing cells in the population of W3110 *ΔompT* pAVE011 cultures **B)** Fluorescence of top 10% sfGFP expressing cells in the population of BL21 pAVE011 cultures **C)** Fluorescence of top 10% sfGFP expressing cells in the population of W3110 *ΔompT* pIAH011 cultures **D)** Fluorescence of top 10% sfGFP expressing cells in the population of BL21 pAVE011 cultures. Measurements for each culture were taken during weekly FACS.

The B background strains have shown a more considerable improvement in the three measured areas (fluorescent fraction of population and median fluorescent values of the entire fluorescent population and the top 10% producers) than the K background strains. The underlying cause is suspected to be genetic and will be investigated further via sequencing. As expected, pIAH011-carrying strains evolved

with different results than strains carrying the original pAVE011 plasmid. This is likely due to the divergent evolutionary paths taken due to pIAH011 inability to lose the promoter via recombination, which is the preferred route of protein production burden reduction in pAVE011-carrying strains.

The differences observed between B and K background strains carrying the original pAVE011 plasmid highlight the varied response two closely related host strains can exhibit when coevolving with the same plasmid. It has been reported before (Hall et al. 2021) that specific genetic conflicts may cause a high burden of plasmid carriage, and when they are resolved, the burden lessens. Such a specific conflict between the pAVEway plasmids and K background *E.coli* could explain why their productivity drastically drops after the first exposure to the inducer and fails to recover over the next several weeks of evolution. The drop in productivity with subsequent recovery in B background strains is consistent with the coevolutionary principle of mutations arising in a specific order, where one mutation is necessary for the next mutation to arise (Bottery et al. 2018).

4.2.3 Phenotypic analysis of evolved populations and their comparison with ancestral counterparts

While the results described in section 4.2.2 highlight the differences between evolutionary outcomes of different genetic background and plasmid combinations, it is crucial to understand these differences on a single clone level.

Using an equation adapted from Leveau & Lindow (2001) to estimate the rate of sfGFP production, a 96-well plate protocol was designed to assess the following features of the clones: overall productivity, overall protein production in the absence of inducer ("leaky expression"), and response to repeated inducer exposure (robustness). For clarity, the production rate is referred to as “culture productivity” and expression in the absence of the inducer as “leaky expression”. It is important to note that the expression from pAVEway vectors in the absence of IPTG is not truly due to a low expression rate from repressed promoter; instead, as described previously in section 3.2.7, it is due to the derepression of the system by complex media components such as tryptone and yeast extract as well as growth by-products in the media present in later stages of culture growth.

Phenotyping data was then used to identify clones that showed a stable change compared to ancestral profiles. These clones were then further investigated to identify underlying genetic changes.

4.2.3.1 The phenotyping protocol

The following protocol has been developed to assess both ancestral and evolved *E. coli* strains. Rather than assessing a mixed population, twelve clones per each ancestral and evolved population were chosen for phenotyping. This is because the phenotype of a mixed population is a sum of genotypes present within the

population, whereas a phenotype of a clonal population can be directly linked to its unique genotype.

For a visual schematic of this protocol, see **Fig. 4.8**. All six populations of each of the four ancestral and evolved strains (1515, 1516, IAH015, IAH016) were streaked on agar plates and grown overnight to obtain single colonies. Twelve of these were then patched onto a new agar plate and incubated again to create a stock plate with twelve clone representatives of each population. The stock plates were stored at 4°C until the 96-well plate assay date. The storage length varied between one and three days. The length of storage at 4°C was minimised to ensure the least amount of variation introduced by it. As evidenced by ancestral data presented in Fig. **4.11** and **4.12**, the response was not dependent on the length of storage (clones from populations 1 and 4 were assayed after one day in storage; clones from populations 2 and 5 after two days in storage; clones from populations 3 and 6 after three days in storage).

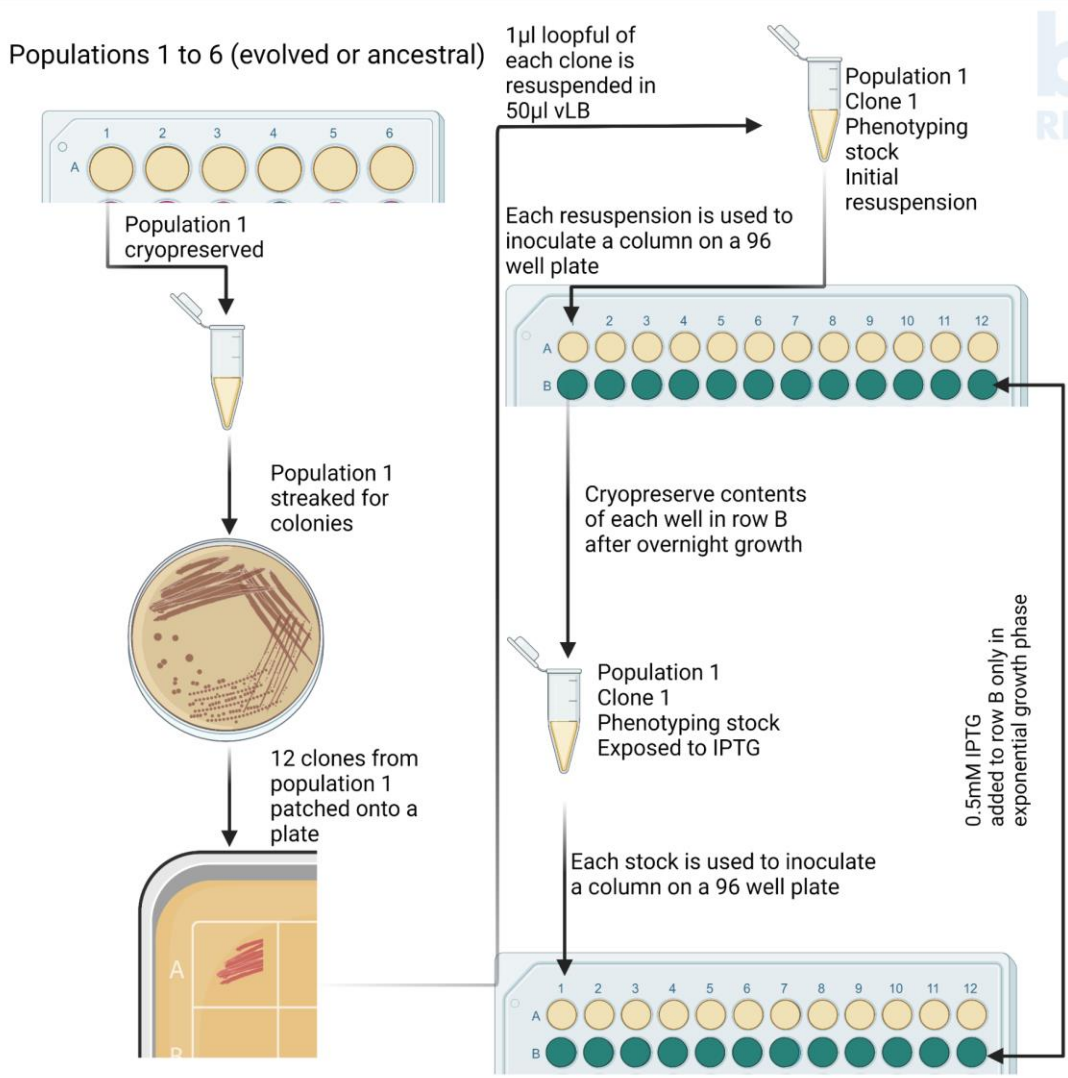


Fig. 4.8 Phenotyping protocol schematic. This protocol was repeated for all 12 clone isolates from each of the six evolved and ancestral populations. All media was supplemented with tetracycline to a final 10 µg/ml concentration. Cryopreservation involved mixing with 40% sterile glycerol in a 1:1 ratio and storing at -70°C. All incubation steps were carried out at 37°C. The 96-well plates were shaken while incubating (200 rpm), and OD_{600} and fluorescence reading was taken every 15 minutes. Illustration created with biorender.com

A single loopful (1µl) of each stock clonal patch was resuspended in 50µl of vLB media to create phenotyping resuspension culture. 2 µl of each culture was used to inoculate 197µl vLB in 2 wells on a 96-well plate and the rest was cryopreserved afterwards for reference. The plate was incubated in a plate reader, shaking. OD_{600} and fluorescence readings were taken every fifteen minutes. In the mid-log phase (usually between 2.5 to 5 hours post inoculation), 1µl of IPTG was added to one of the wells, while 1µl of sterile MilliQ water was added to the other well. The plate was then returned to the plate reader for overnight incubation. This is further described as the "first induction" or "first IPTG exposure" protocol.

The next day the induced cultures from the plate were cryopreserved. These were thawed later, and 2µl of each culture was used to inoculate 197µl of vLB media in 2 wells of a 96-well plate. The plate was incubated in a plate reader with orbital shaking. OD_{600} and fluorescence readings were taken every fifteen minutes. In the mid-log phase (usually between 2.5 to 5 hours post inoculation), 1µl of IPTG was added to one of the wells, while 1µl of sterile MilliQ water was added to the other well. The plate was then returned to the plate reader for overnight incubation. This is further described as the "second induction" or "second IPTG exposure" protocol.

4.2.3.2 Calculation of culture productivity

The average promoter activity calculation of the various strains post-induction involved some mathematical transformation of the raw data. An accurate comparison of the promoter activity between strains was possible due to applying an altered equation described in (Leveau & Lindow 2001). For detailed description of this equation, see methods.

In this study, the P value of calculated promoter activity is referred to as "estimated culture productivity", which is more accurate than "promoter activity" in the study context. For a visual representation of the steps in calculating productivity see **Fig. 4.9.**

Immediately before inoculation, the media baseline values were measured and averaged. Then, these baseline values were subtracted from the later data points.

For each time point of the phenotyping experiment, two values were obtained - OD_{600} and fluorescence. The $\ln(OD_{600})$ values were calculated and plotted against time **(Fig. 4.9A)** to obtain growth rate μ values for each 15-minute segment of the timeline. Fluorescence values were plotted on a separate graph against OD_{600} values **(Fig. 4.9B)** to obtain fluorescence steady-state values for each 15-minute segment of the timeline. Each 15-minute segment of the experimental timeline had two values associated with it - its specific growth rate and fluorescence steady-state constant. The following time segments were excluded from further analysis **(Fig. 4.9C)**:

- If its growth rate was negative
- If its fluorescence steady-state value was negative
- If both were negative.

This was to prevent negative values of the calculated culture productivity P (as culture cannot be producing a negative amount of protein) and falsely positive P values due to negative growth rate and negative fluorescence steady state being multiplied by each other. For each of the remaining segments, the culture productivity was calculated using the adapted equation **(as shown in Fig. 4.9C)**. These productivity values were then plotted against time **(Fig. 4.9D)**, and the area under the curve (AUC) was calculated for the curve encompassing all timepoints from induction until 15 hours

later. AUC values were then compared to assess the productivity and induction tolerance of the evolved strains.

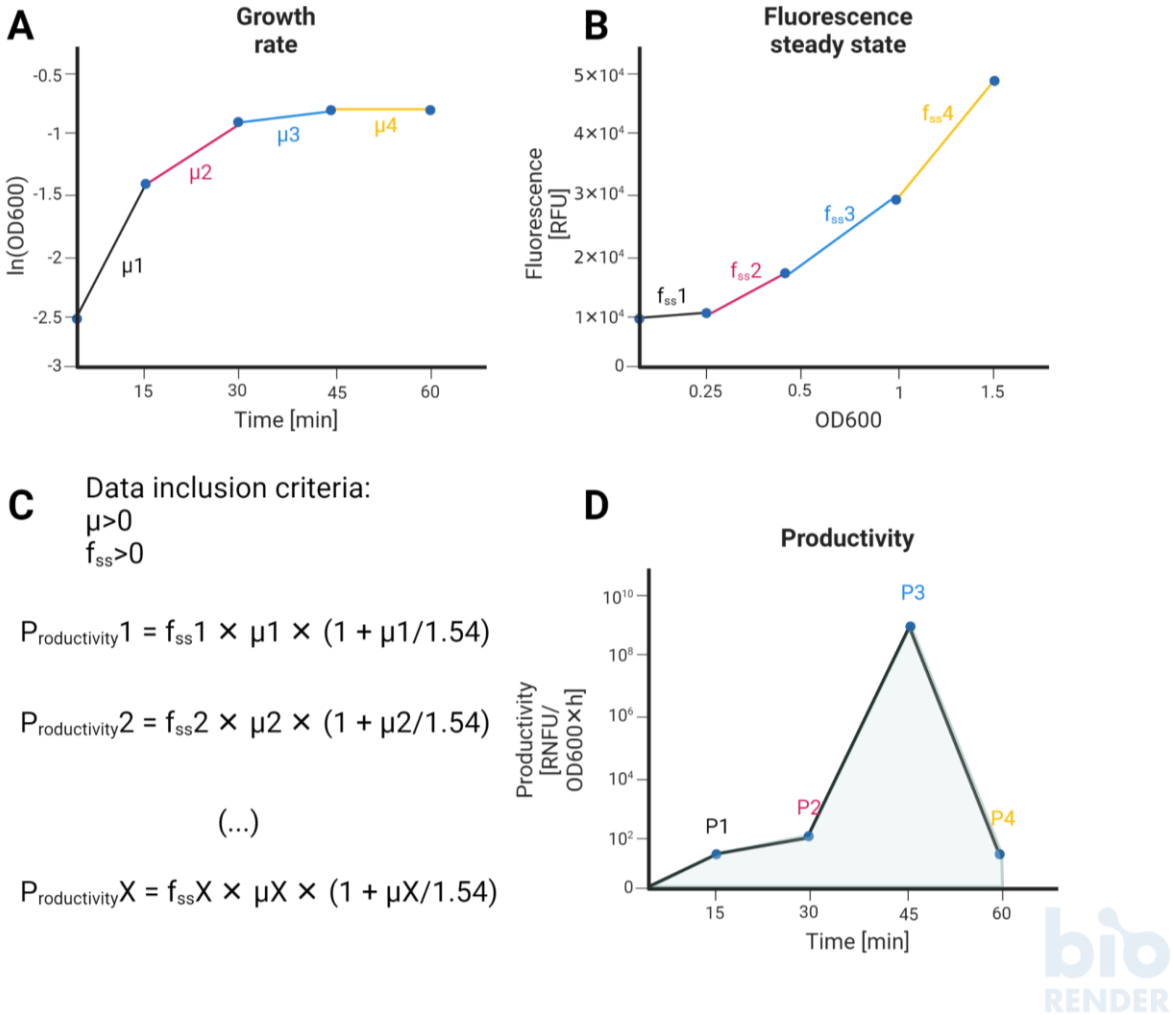


Fig. 4.9 A visual representation of culture productivity calculations using an adapted equation (Leveau & Lindow 2001)**.** All calculations in steps **A, B and C** were performed in Excel. Data exclusion was performed using IF(AND(...)) function. Culture productivity data was further analysed in GraphPad Prism. Illustration created with biorender.com. **A)** Natural logarithm of OD600 values for each timepoint was calculated and plotted against time to estimate growth rates μ for each 15-minute segment of the experimental timeline. **B)** Fluorescence values were plotted against associated OD_{600} values to estimate fluorescence steady-state f_{ss} for each 15-minute segment of the experimental timeline. **C)** After excluding time segments which did not meet the inclusion criteria, culture productivity P was calculated for each 15 min segment of the experimental timeline. **D)** The calculated P values were plotted against time. Note that the unit (**R**elative **N**on-**F**luorescence **U**nits per OD_{600} per hour) indicates

the amount of immature, non-fluorescent sfGFP produced per each OD_{600} unit of the culture per hour. The area under the curve was then calculated for the curve segment of 15 hours post-induction for each culture.

4.3.2.3 Ancestral pAVEway plasmid-carrying strains characteristics

To assess which evolved lines to investigate further by genetic sequencing, a protocol to quantify improvement of the evolved strains when compared to ancestors was developed where the following factors were considered:

1. Productivity in the absence of inducer
2. Productivity in the presence of inducer
3. Stability of the response observed during the first induction.

First, the initial response to standard inducer concentration (0.5 mM IPTG) had to be assessed. Six ancestral populations were first characterised by employing the phenotyping protocol described in **Fig. 4.8** to establish baseline responses. Twelve biological replicate cultures originating from each population were grown in technical duplicate in a 96-well plate overnight, and growth and fluorescence data were recorded in 15-minute intervals. One of the technical replicates of each culture was also induced in the mid-log phase. The recorded OD_{600} and fluorescence values were then used to calculate culture productivity (see **Fig. 4.9** for method details) of induced and uninduced samples. These productivity values were plotted against time, and the area under the curve (AUC) of the resulting charts was calculated. In order to ensure the values were comparable between experiments, only the data collected within the first 15 hours after induction was considered in this calculation. A representative dataset is presented in **Fig 4.10** and showcases one of the trends observed during the phenotyping experiment predominantly in BL21 background strains. Upon 1st exposure to the inducer, ancestral 1516 populations produce high quantities of sfGFP **(Fig. 4.10A).** This productivity is the highest just after induction, shows a downward

trend after the first hour, and drops to levels comparable to those of an uninduced culture approximately 15 hours post induction. When this population is then exposed to the inducer for the second time **(Fig. 4.10C)**, the production of sfGFP is considerably lower than during the first exposure to the inducer **(Fig. 4.10C).** This is in contrast to the trends observed in the evolved population **(Fig. 4.10B and D)**. During 1st exposure to the inducer, the evolved population maintains its productivity at a high and stable level for over 20 hours **(Fig 4.10B).** During the second exposure to the inducer, the same culture shows improved productivity when compared to the ancestral population responding to the repeated induction stress.

A geometric mean and its 95% confidence interval (CI) was then calculated from 12 AUC values for each ancestral population to provide a reference for comparisons with the evolved strains **(Fig. 4.11 and 4.12)**. This data also provided a reference value for how productive the populations were in the absence of induction (the expression non-specifically induced by complex media components.

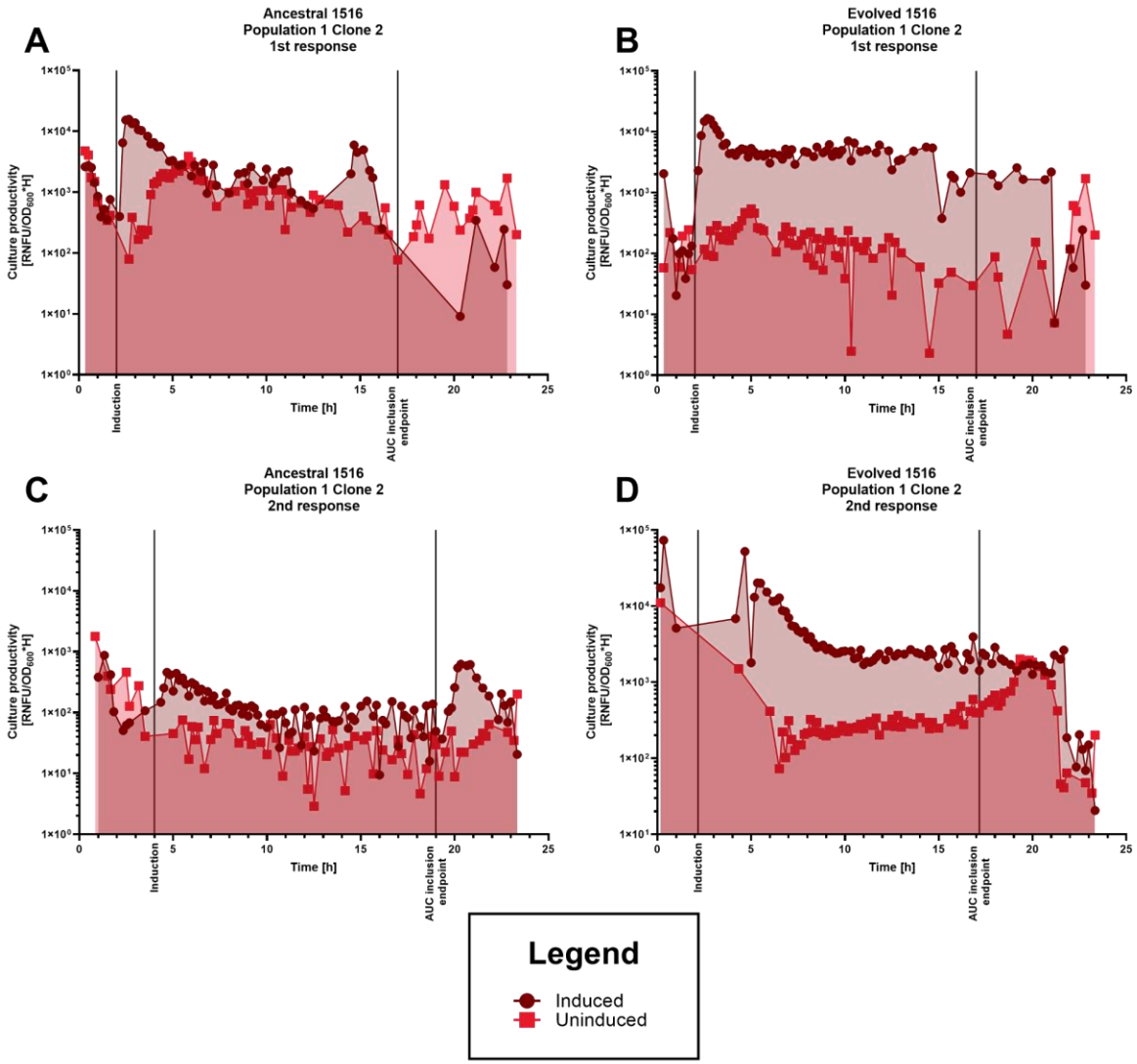


Fig. 4.10 An example of trends observed in culture productivity over time during phenotyping. Each chart contains both induced and uninduced data for the same clone isolated from one of the 6 populations. **A)** Ancestral clone 2 isolated from population 1; first response to inducer **B)** Evolved clone 2 isolated from population 1; first response to inducer **C)** Ancestral clone 2 isolated from population 1; second response to inducer **D)** Evolved clone 2 isolated from population 1; second response to inducer

Vertical lines were drawn at the induction time point, as well as 15h post induction.

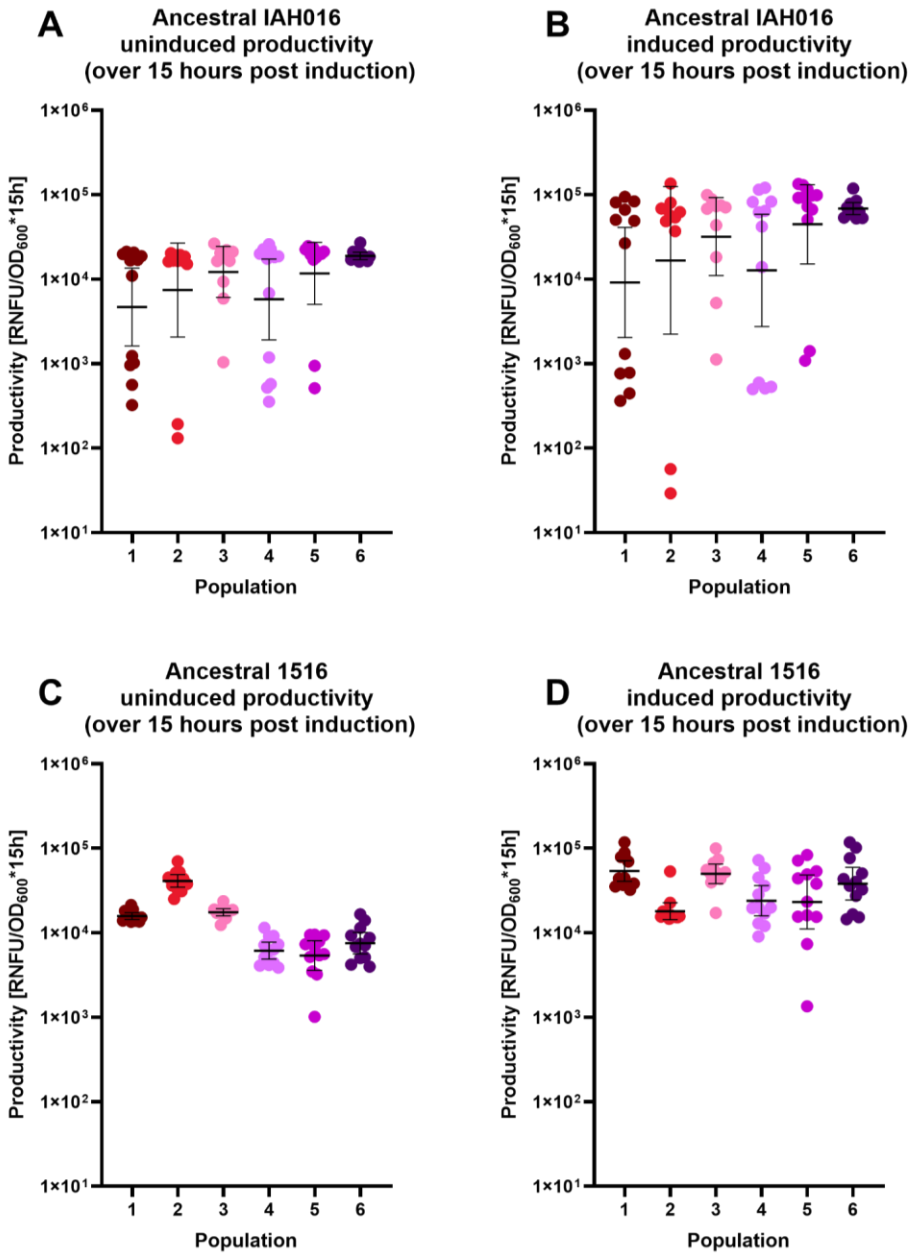


Fig. 4.11 The cumulative culture productivity calculated over 15 hours after induction for ancestral IAH016 (A, B) and 1516 (C, D) strains. A) and **C)** Culture productivity of 12 clones isolated from each population in the absence of inducer (IPTG) facilitated by non-specific induction through complex media elements such as tryptone and yeast extract. **B)** and **D)** Culture productivity of 12 clones isolated from each population induced by the addition of IPTG in the mid-log growth phase (final concentration 0.5 mM). Each point represents a clone. The purple bars represent geometric means with their 95% CI.[1]

[1] The strain designations used in the figures within this chapter do not match the genetic sequencing data obtained later. This is discussed further in Chapter 5.

IAH016 and 1516 ancestral populations varied in productivity without induction **(Fig. 4.11A, C)**. The geometric mean of productivity of IAH016 population 6 was appreciably higher than that of populations 1-5 **(Fig. 4.11A)**. One of the reasons for this observation is that some - between 17 to almost 42% - of the of the 12 clones isolated to characterise each population from 1 to 5 did not show a significant increase in productivity even after induction **(Fig. 4.11B)**, while all clones isolated from population 6 did. The same trend is observed in all IAH015 populations **(Fig. 4.12A, B)**. This failure to respond to inducer stimuli is unlikely to be caused by a complete plasmid loss (as media was supplemented with tetracycline). Instead, it represents naturally occurring population heterogeneity resulting from either random mutations altering expression control elements (repressor, operator, promoter sequences) or plasmid copy number variation between the clonal cultures. pAVEway plasmids are medium to high copy number plasmids with no active partitioning system (ColE1 derivative). Therefore, some plasmid copy number variation is expected due to the random distribution of plasmid copies into the daughter cells (Reyes-Lamothe et al. 2014). Additionally, some recent studies have shown that plasmids are not distributed within the cell in a truly random fashion. Evidence from microscopy shows that high plasmid copy number (PCN) plasmids can exhibit a hybrid behaviour between randomised diffusion and clustering (Wang et al. 2016) This can enhance the number of lower PCN daughter cells. Moreover, in the case of plasmids encoding costly recombinant protein production machinery, these daughter cells will then have a significant fitness advantage against high PCN daughter cells.

In 1516 ancestral populations, the expression of sfGFP in the absence of the inducer is higher for populations 1, 2 and 3 than 4, 5 and 6. This corresponds with proportionately higher levels of expression when induced. This clustering of data can

be explained by the differences in ancestral phenotypes present in two cryostock tubes used to inoculate populations 1-3 and 4-6. While IAH016 and IAH015 populations show a lower degree of non-specific induction than strains 1515 and 1516, they also produce a lower amount of sfGFP when induced with the same amount of IPTG **(Fig. 4.11, 4.12)**. This may be due to the structure of pIAH011 plasmid operators, which are non-homologous palindromes, unlike the homologous palindromes present in pAVE011 plasmids.

There are also more non-responsive clones in pIAH011-carrying populations than pAVE011-carrying populations **(Fig. 4.11, 4.12)**. suggesting that the burden of protein production may be higher for this plasmid-strain combination, resulting in higher fitness cost and more frequent rates of non- and low-producing clones arising. Alternatively, the mechanism used to escape the protein production burden depends on the plasmid variant carried by the host. In pAVE011 plasmids, the promoter region can be deleted via homologous recombination between the palindromic operators, resulting in a lower productivity. Meanwhile, the pIAH011 plasmids do not undergo homologous recombination at high rates. Therefore, plasmid copy number variation during cell division may be providing a more significant evolutionary advantage to the daughter cells with lowered PCN than it does in pAVE011 carrying strains. Lowered PCN would also correspond with lowered productivity and lowered responsiveness upon induction.

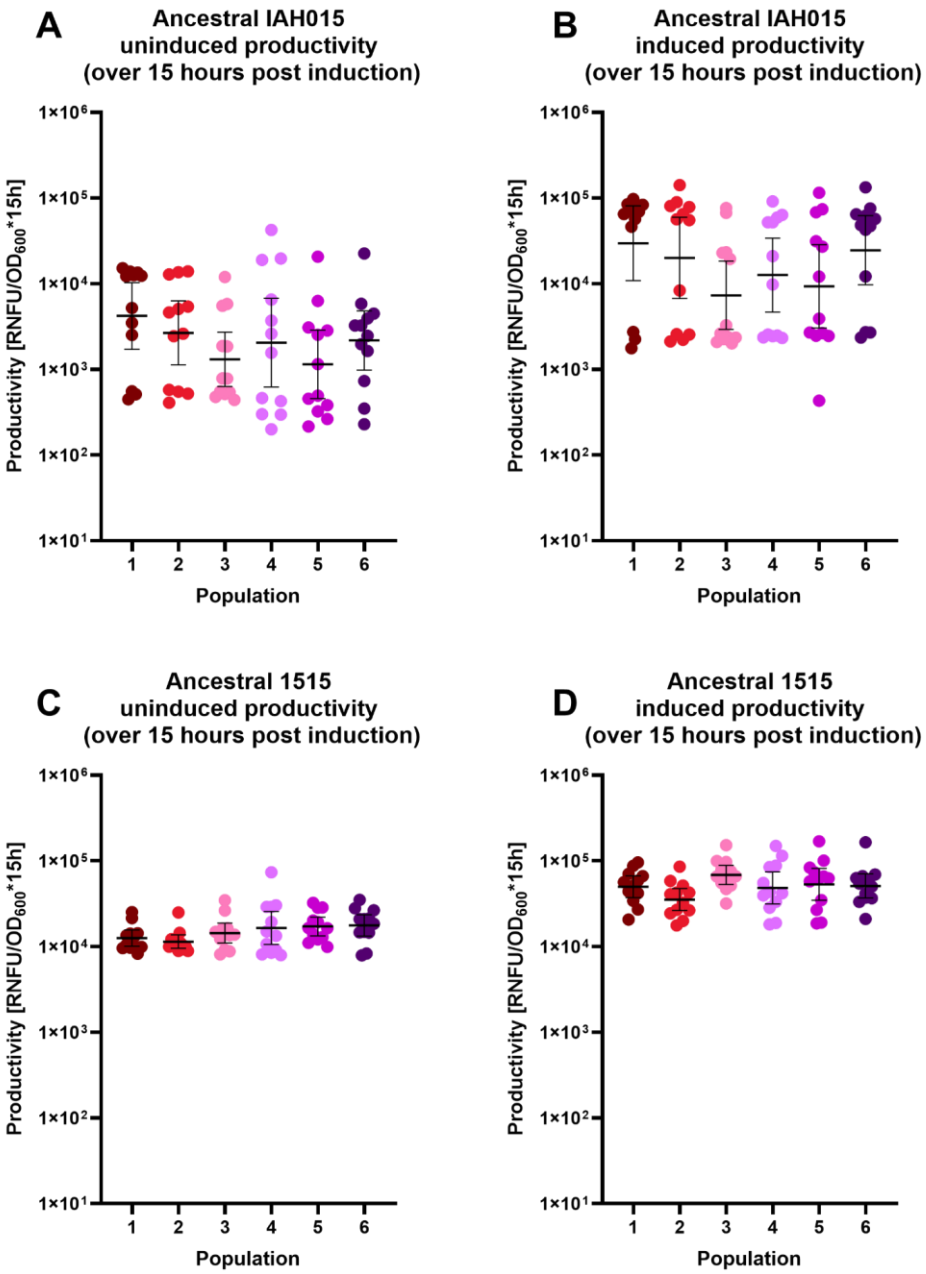


Fig. 4.12 The cumulative culture productivity calculated over 15 hours after induction for ancestral IAH015 (A, B) and 1515 (C, D) strains. A) and **C)** Culture productivity of 12 clones isolated from each population in the absence of inducer (IPTG) facilitated by non-specific induction through complex media elements such as tryptone and yeast extract. **B)** and **D)** Culture productivity of 12 clones isolated from each population induced by the addition of IPTG in the mid-log growth phase (final concentration 0.5 mM). Each point represents a clone. The purple bars represent geometric means with their 95% CI.[2]

[2] The strain designations used in the figures within this chapter do not match the genetic sequencing data obtained later. This is discussed further in Chapter 5.

The observed bimodal distribution of productivity in the ancestral strains of pAVE011 and pIAH011 carrying strains has implications for the geometric means and 95% confidence intervals calculated and used in later comparisons. There are alternative approaches which could be utilised to calculate the mean productivity values. Firstly, the clones which failed to respond to 0.5 mM IPTG induction could be removed from the dataset as outliers. This could however lead to falsely elevated means for the entire population, and it would fail to capture the mixed nature of the ancestral populations, where responses to induction vary between clones of the same population. This is an important trait which distinguishes some ancestral populations from the evolved ones (bimodality of response is present in figures **4.11** and **4.12** in contrast to data clustering seen in **Fig. 4.15C, E, F; 4.16, 4.17A** and **B** and **4.18A, B** and **C)** Another approach would be to calculate one ancestral mean for every strain by incorporating all the data from all six populations. This approach also has its challenges. Calculating a single mean for all ancestral populations would disregard the variation between the responses of different populations. Therefore, the best approach for this study was to consider each ancestral population separately and compare it only to the clones isolated from the evolved population directly descended from it; regardless of the widely spread 95% CI values.

The second important factor to consider was the stability of the response observed when first exposed to IPTG. It has been previously shown that ancestral pAVE011-carrying strains do not maintain the same level of response when induced repeatedly **(Fig. 4.10)**, and it was crucial to establish whether the same is observed in the evolved strains. Additionally, the stability of the induction response could suggest

that the observed phenotypic adaptation was due to an underlying genetic change rather than a transient change, such as sustained changes in specific stress response gene expression levels. Strain stability estimation was carried out by dividing the AUC values during the second response by those during the first response (**Fig. 4.13, 4.14**). In an ideally stable population the resulting ratio would be 1.

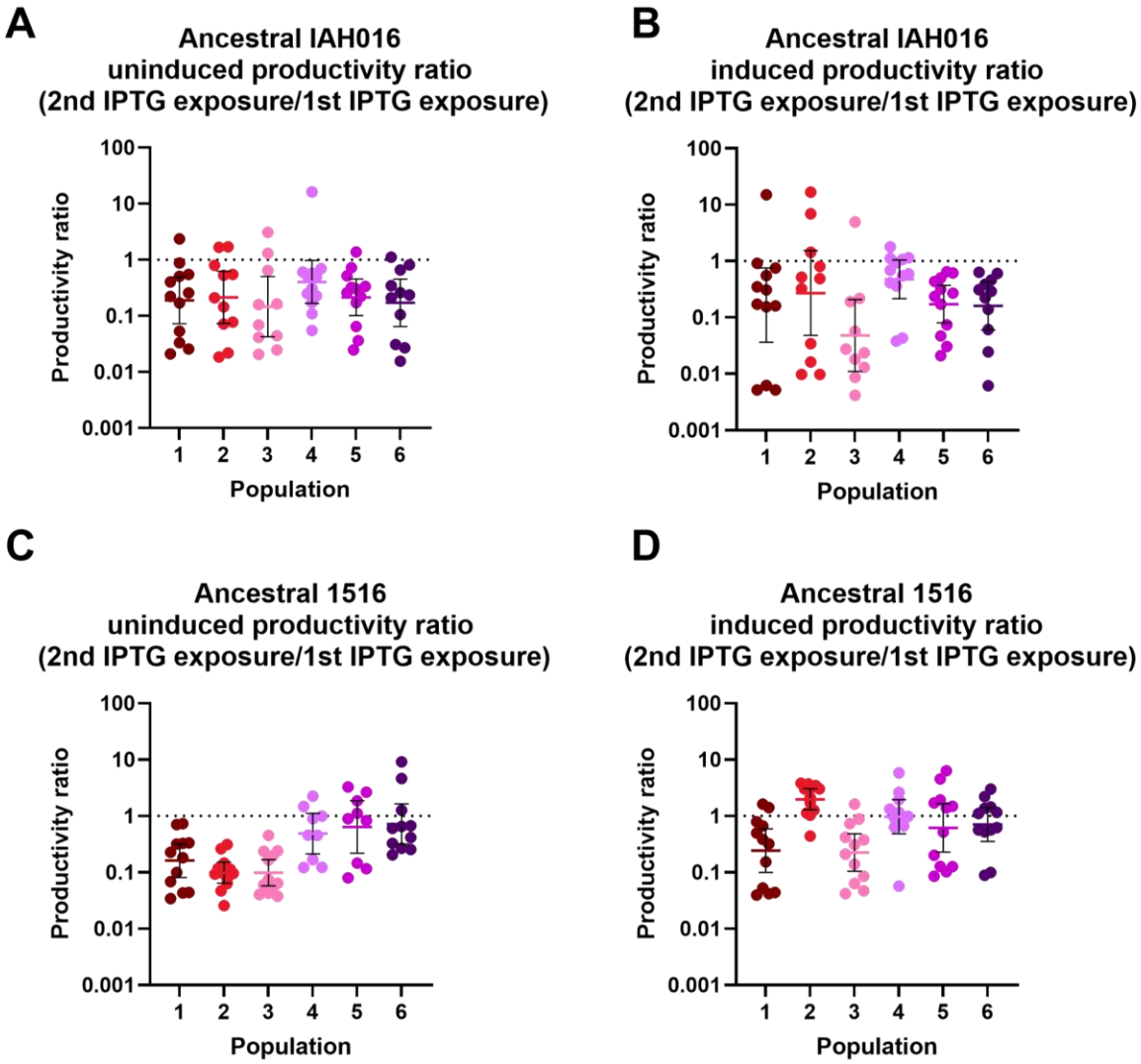


Fig. 4.13 The culture productivity ratios of IAH016 and 1516 strains. The ratios are calculated as the AUC values of the 15-hour post-induction period during the second IPTG exposure divided by matched values from the first IPTG exposure experiment. If the response in both experiments is the same (stable), this value is expected to be 1 (represented by the dotted line at Y=1). Values below 1 indicate lower productivity during the second induction. Productivity ratios were calculated for 12 representative clones from each population, represented by individual data points. **A)** and **B)** Uninduced and induced ancestral IAH016 productivity rates. **C)** and **D)** Uninduced and induced ancestral 1516 productivity rates. Lines and error bars represent the geometric mean and its 95% CI.[3]

[3] The strain designations used in the figures within this chapter do not match the genetic sequencing data obtained later. This is discussed further in Chapter 5.

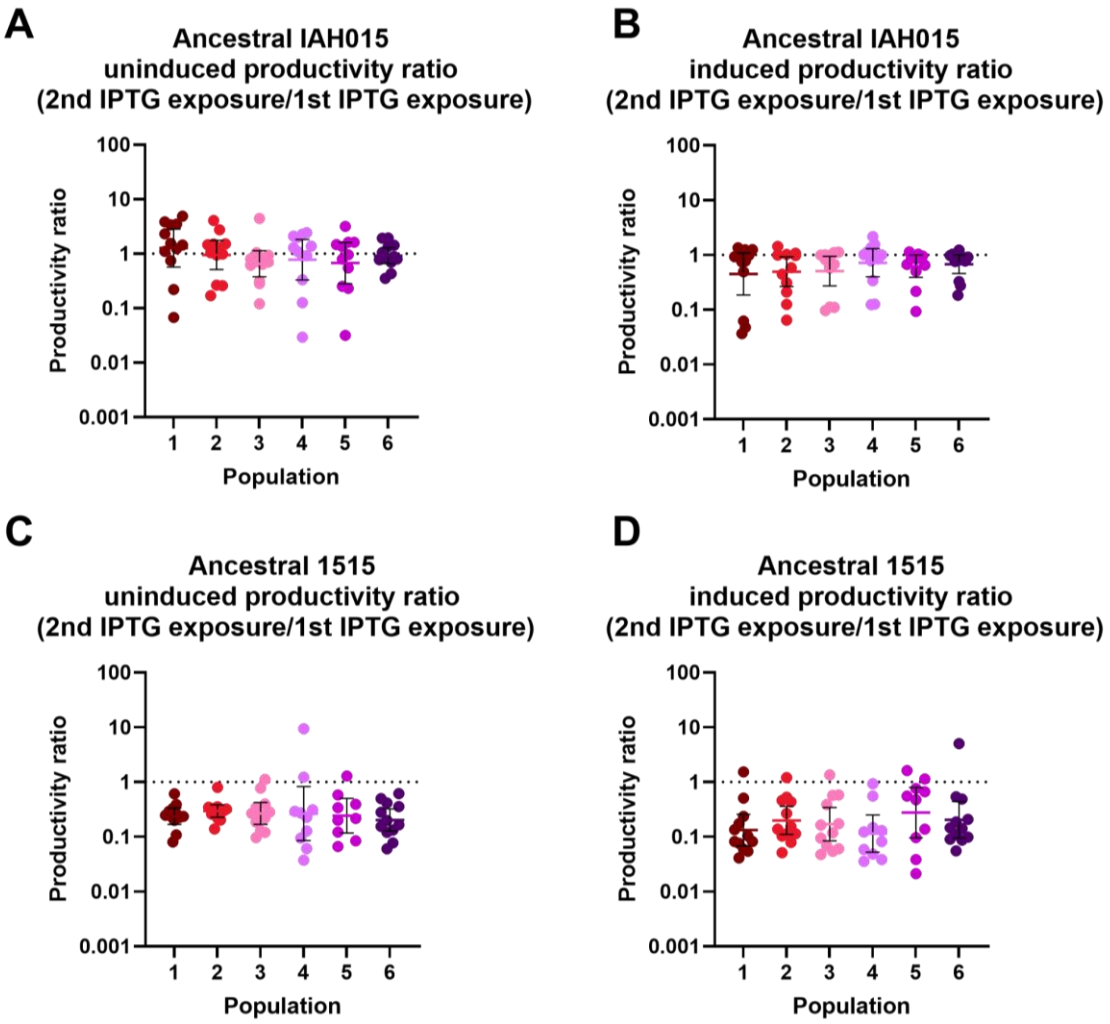


Fig. 4.14 The culture productivity ratios of IAH015 and 1515 strains. The ratios are calculated as the AUC values of the 15-hour post-induction period during the second IPTG exposure divided by matched values from the first IPTG exposure experiment. If the response in both experiments is the same (stable), this value is expected to be 1 (represented by the dotted line at Y=1). Values below 1 indicate lower productivity during the second induction. Productivity ratios were calculated for 12 representative clones from each population, represented by individual data points. **A)** and **B)** Uninduced and induced ancestral IAH015 productivity rates. **C)** and **D)** Uninduced and induced ancestral 1515 productivity rates. Lines and error bars represent the geometric mean and its 95% CI.[4]

[4] The strain designations used in the figures within this chapter do not match the genetic sequencing data obtained later. This is discussed further in Chapter 5.

In IAH016 ancestral populations, all mean uninduced ratios are below 1, indicating that the non-specific induction by media components is less pronounced during second-time exposure to the inducer (**Fig. 4.13A**). This could suggest an adaptive change reducing the metabolic burden of plasmid carriage happening rapidly. However, if this is the case, those same changes are causing an impaired response to second-time induction, as all induced productivity ratios are also below 1 (**Fig. 4.13B**). The same trend is observed in the strain 1515 **(Fig. 4.14C, D).**

In 1516 ancestral populations 1, 2 and 3, most of the uninduced and induced ratio geometric means are below 1, while those of populations 4, 5 and 6 are above 1 (**Fig. 4.13C**). This correlates with the populations 4, 5 and 6 showing lower rates of non-specific induction during first exposure to IPTG (**Fig. 4.11C**). It provides further evidence that lower production of sfGFP in the absence of the inducer provides a fitness advantage when competing with high producer clones.

The most stable ancestral strain is IAH015, a K background host carrying the altered pIAH011 plasmid **(Fig. 4.14A, B).** Similarly to the other three strains, it shows reduced response during the second induction, however this effect is the least pronounced in this strain. This highlights the relevance of specific strain/plasmid combinations on strain fitness.

While continuous fermentation is gaining popularity in the recombinant protein production industry due to its cost efficiency (Li et al. 2014), the instability of pAVEway plasmid-carrying ancestral strains is a significant factor limiting the potential of the pAVEway platform being used in such protocols. Furthermore, the same trend of reduced responsiveness during second induction is observed in all four tested strains (with varying intensity), which indicates that the issue is unlikely to be caused solely by plasmid-strain-specific interactions. Instead, repeated exposure to IPTG acts as a

selective pressure, enriching the populations in low- or non-producer clones. This is because avoiding the protein expression from a powerful promoter in pAVEway plasmids is the easiest way to reduce the metabolic burden and gain a fitness advantage. A similar adaptive mechanism to overexpression of sfGFP in *E. coli* has been observed in another study (James et al. 2021), where adaptive mutations were most commonly found in regions responsible for the recombinant protein production control (T7 polymerase, promoter/operator region).

4.3.2.4 Phenotypic characterisation of the evolved strains

4.3.2.4.1 First induction response in several clones isolated from the evolved strains is improved compared to their ancestors

For the evolved clones to be considered “improved”, their phenotype characteristics must show enhanced usefulness for industrial protein production. The first useful property is lower uninduced productivity han that of the ancestral population. This will improve the strain’s ability to produce heterologous proteins harmful to *E. coli* and more generally decrease the burden on the cell. In such cases, it is crucial for the system to remain repressed in the absence of an inducer to allow for maximal biomass accumulation before the production of protein is induced, resulting in death in host cells (Rosano et al. 2019). Secondly, their induced productivity should be higher than that of the ancestral population. This will contribute to an increase in yield and, therefore, higher cost-effectiveness of production processes. Finally, the response to the inducer should be more stable than that of the ancestors, prolonging the production window of the culture and allowing for continuous processes to be adapted.

Using the same phenotyping protocol as for the ancestors, the evolved IAH016, 1516, IAH015 and 1515 populations were characterised (six of each).

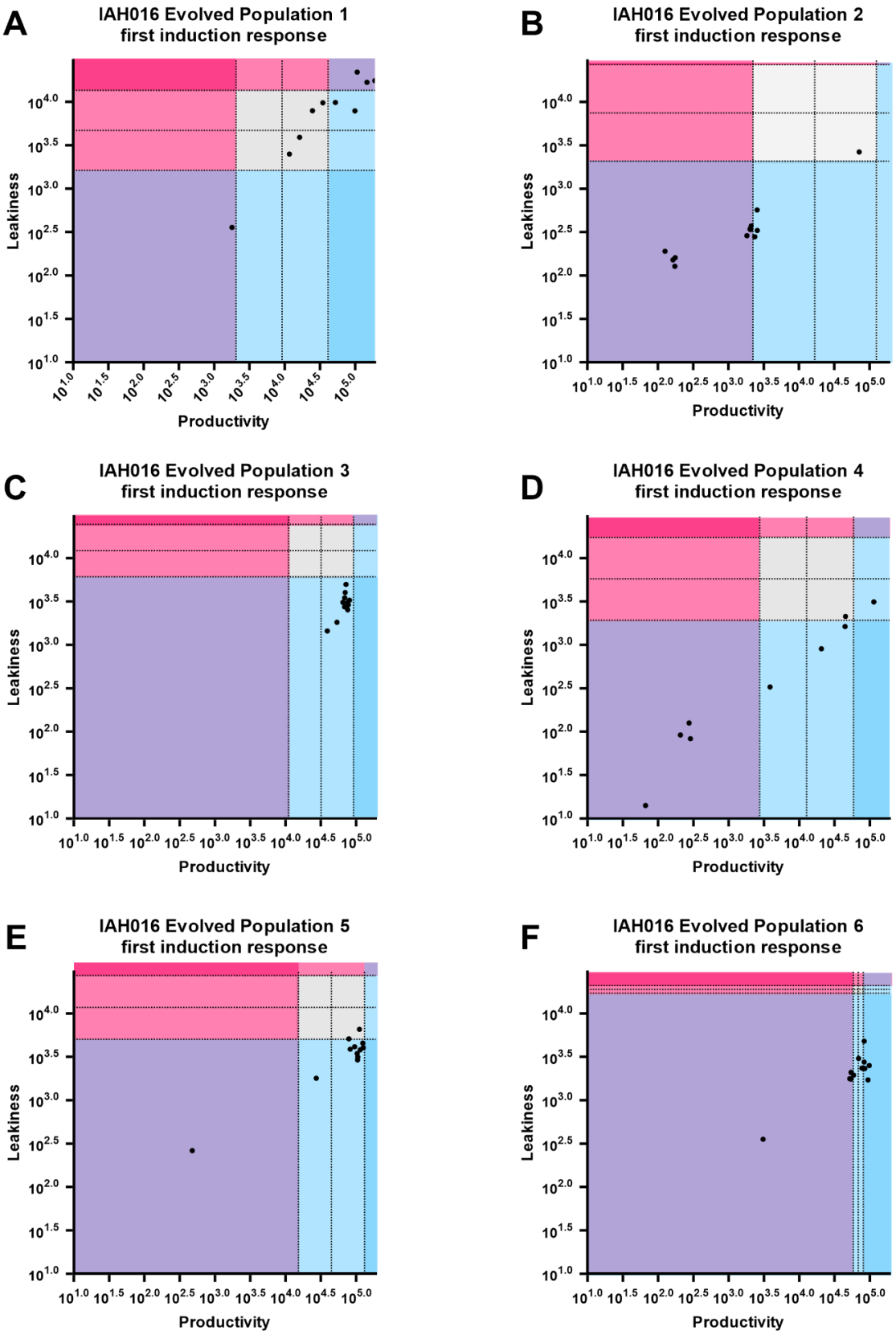


Fig. 4.15 The first induction response of six evolved IAH016 populations. The productivity was calculated as the AUC of the first 15 hours post-induction. The solid lines represent geometric means of uninduced and induced productivity of the relevant

ancestral population, and the dotted lines represent 95% CI of those means. The colours of the boxes represent the phenotype relative to ancestors assessed by two parameters (induced and uninduced productivity). **Grey** - ancestral first induction response; **purple** - one parameter is better while the other is worse; **dark magenta -** both parameters are worse; **dark blue -** both parameters are better; **light magenta** - one parameter is worse while the other did not change; **light blue** - one parameter is better while the other did not change.[5]

Several phenotype profiles are observed within the populations of the IAH016 strain (**Fig. 4.15**), and these profiles vary in distribution between populations. For example, in populations 3, 5 and 6 (**Fig. 4.15C, E, F**) the majority of the isolated clones are captured within the blue boxes, signifying improvement in at least one of the two assessed parameters. This means either their productivity in the absence of the inducer is lower than ancestral, their induced productivity is higher than ancestral, or both. In contrast, the phenotype distribution in populations 1, 2 and 4 (**Fig. 4.15A, B, D**) is much different, with the majority of clones showing either no change (**Fig. 4.15A**) or an improvement with a trade-off profile whereby either uninduced or induced productivity is improved at the cost of the other worsening when compared to the ancestral response (**Fig. 4.15A, B, D**). These findings suggest that each of the six populations has followed a distinct evolutionary path to achieve improved fitness.

[5] The strain designations used in the figures within this chapter do not match the genetic sequencing data obtained later. This is discussed further in Chapter 5.

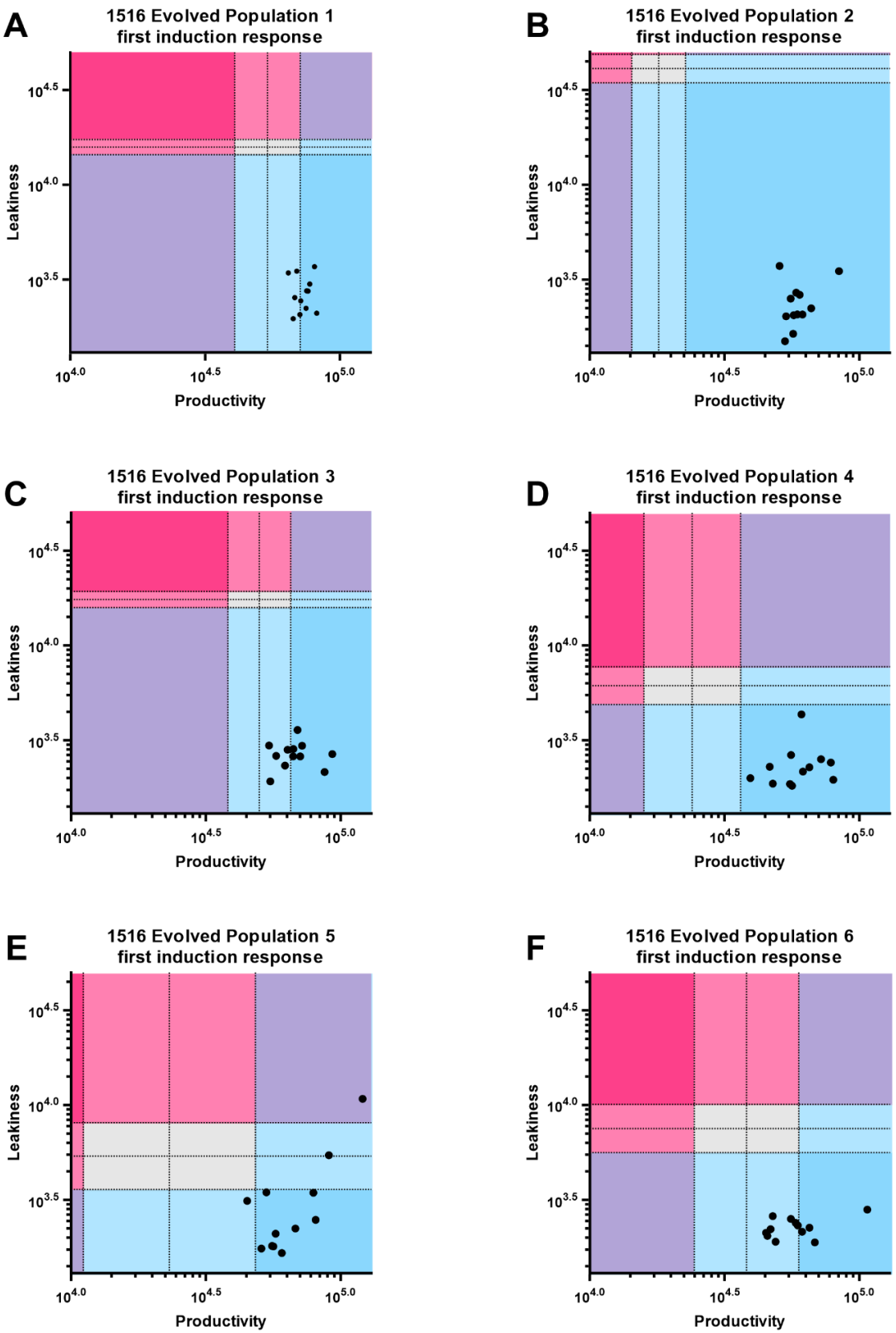


Fig. 4.16 The first induction response of six evolved 1516 populations. The productivity was calculated as the AUC of the first 15 hours post-induction.The solid lines represent geometric means of uninduced and induced productivity of the relevant

ancestral population, and the dotted lines represent 95% CI of those means. The colours of the boxes represent the phenotype relative to ancestors assessed by two parameters (induced and uninduced productivity). **Grey** - ancestral first induction response; **purple** - one parameter is better while the other is worse; **dark magenta -** both parameters are worse; **dark blue -** both parameters are better; **light magenta** - one parameter is worse while the other did not change; **light blue** - one parameter is better while the other did not change[6]

A similar trend can be observed in the populations of the 1516 strain (**Fig. 4.16**), where although most clones fall within the blue boxes showing an improvement in at least one of the parameters when compared to their ancestors, the degree of this change varies between populations, with the most prominent degree of difference observed in populations 1, 2 and 3 (**Fig. 4.16 A, B, C**). When considered in the context of the results shown in **Fig. 4.11C, D** and **Fig. 4.13C, D**, it is plausible that the first necessary step in the evolution of these strains was to reduce the burden of sfGFP production in the absence of the inducer; this change could have already happened in half of the ancestral populations (4, 5, 6). Therefore, while the observed phenotype shift in populations 4, 5 and 6 (**Fig. 4.16D, E, F**) is smaller than that of populations 4, 5 and 6, it may be caused by the lower amount of new adaptations needed to achieve fitness similar to the evolved populations 1, 2 and 3.

[6] The strain designations used in the figures within this chapter do not match the genetic sequencing data obtained later. This is discussed further in Chapter 5.

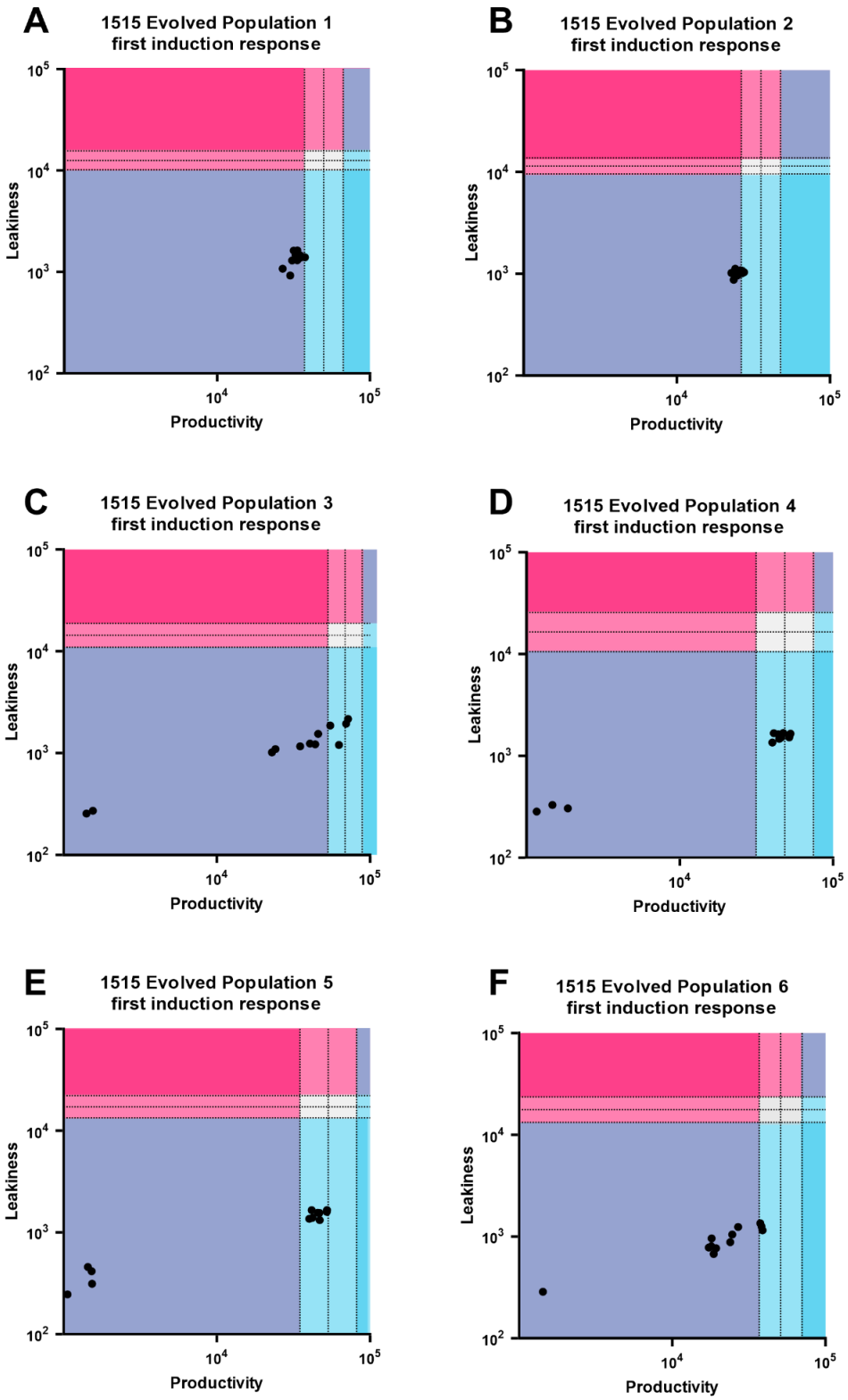


Fig. 4.17 The first induction response of six evolved 1515 populations. The productivity was calculated as the AUC of the first 15 hours post-induction.The solid lines represent geometric means of uninduced and induced productivity of the relevant ancestral population, and the dotted lines represent 95% CI of those means. The

colours of the boxes represent the phenotype relative to ancestors assessed by two parameters (induced and uninduced productivity). **Grey** - ancestral first induction response; **purple** - one parameter is better while the other is worse; **dark magenta -** both parameters are worse; **dark blue -** both parameters are better; **light magenta** - one parameter is worse while the other did not change; **light blue** - one parameter is better while the other did not change[7]

A different evolutionary pattern can be observed in the evolved populations of K background strains 1515 **(Fig. 4.17)** and IAH015 **(Fig. 4.18).** Majority of the evolved clones fall within the purple chart areas, signifying an evolutionary trade-off - while leakiness decreased, so did the productivity **(Fig. 4.17A, B, C, F; Fig. 4.18D, E, F).** This is the expected evolutionary outcome in absence of any selection factors, and can be caused by lower PCN or mutations in the promoter region of the plasmid. However, within each of the 1515 strain populations dominated by this phenotype, there is also a fraction of population showing decreased leakiness and productivity unchanged compared to the ancestors. The population productivity measured over 7 weeks of the evolutionary experiment **(Fig. 4.6A)** revealed an initial drop in population fluorescence within the first week. K background strains (1515 and IAH015) showed the smallest improvement in productivity over the following weeks out of all 4 strains tested. Therefore, the mixed phenotype populations observed later **(Fig. 4.17)** are likely a representation of the evolutionary process still taking place. While some clones show reduced productivity and leakiness consistent with early adaptation to protein overproduction burden, a portion of each population started to recover its productivity due to selection pressure imposed by weekly FACS. A similar trend can be observed

[7] The strain designations used in the figures within this chapter do not match the genetic sequencing data obtained later. This is discussed further in Chapter 5.

in IAH015 populations, suggesting that the K background *E. coli* rate of adaptation may be slower than that of B background *E. coli*. This is supported by the fact that genetic components can influence the mutation rates in *E. coli* (Csörgő et al. 2012) and *E. coli* W3110 and BL21 genomes are significantly different from each other.

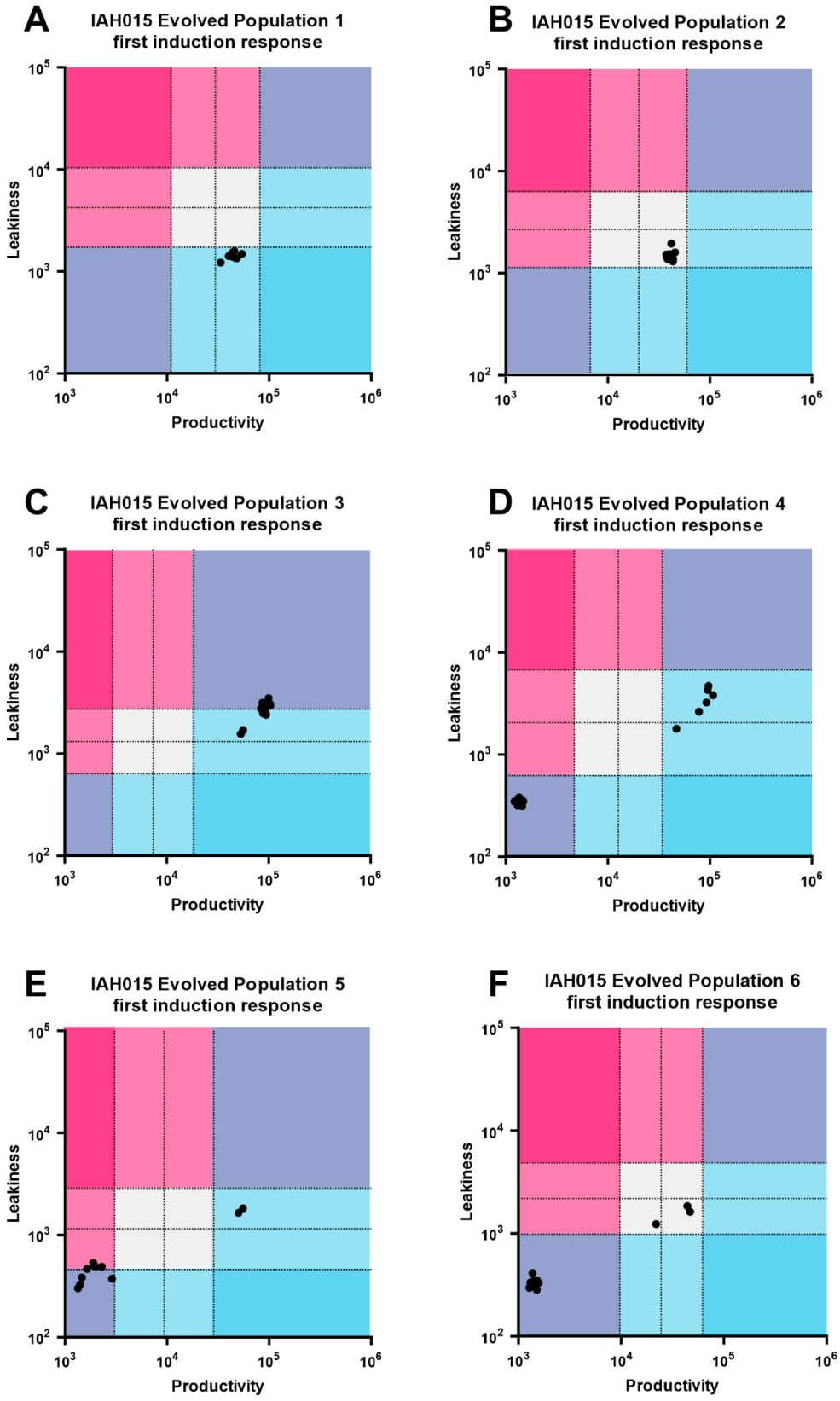


Fig. 4.18 The first induction response of six evolved IAH015 populations. The productivity was calculated as the AUC of the first 15 hours post-induction.The solid

lines represent geometric means of uninduced and induced productivity of the relevant ancestral population, and the dotted lines represent 95% CI of those means. The colours of the boxes represent the phenotype relative to ancestors assessed by two parameters (induced and uninduced productivity). **Grey** - ancestral first induction response; **purple** - one parameter is better while the other is worse; **dark magenta -** both parameters are worse; **dark blue -** both parameters are better; **light magenta** - one parameter is worse while the other did not change; **light blue** - one parameter is better while the other did not change.[8]

[8] The strain designations used in the figures within this chapter do not match the genetic sequencing data obtained later. This is discussed further in Chapter 5.

4.3.2.4.2 Evolved strain stability is improved compared to that of the ancestors

In order to determine whether the change in response to IPTG when compared to ancestors was stable, the strain productivity ratios were calculated. This involved dividing the AUC response recorded during the second induction experiment by that of the first. If the response during the first and second induction is the same, this ratio is expected to be 1. However, because these are measurements of biological replicates, some degree of variation is to be expected. Therefore, the geometric means of induced and uninduced productivity ratios and their 95% CI were calculated for all ancestral populations. The calculated variation was then applied to the number 1 to establish reasonable boundaries for the ideally stable, hypothetical population.

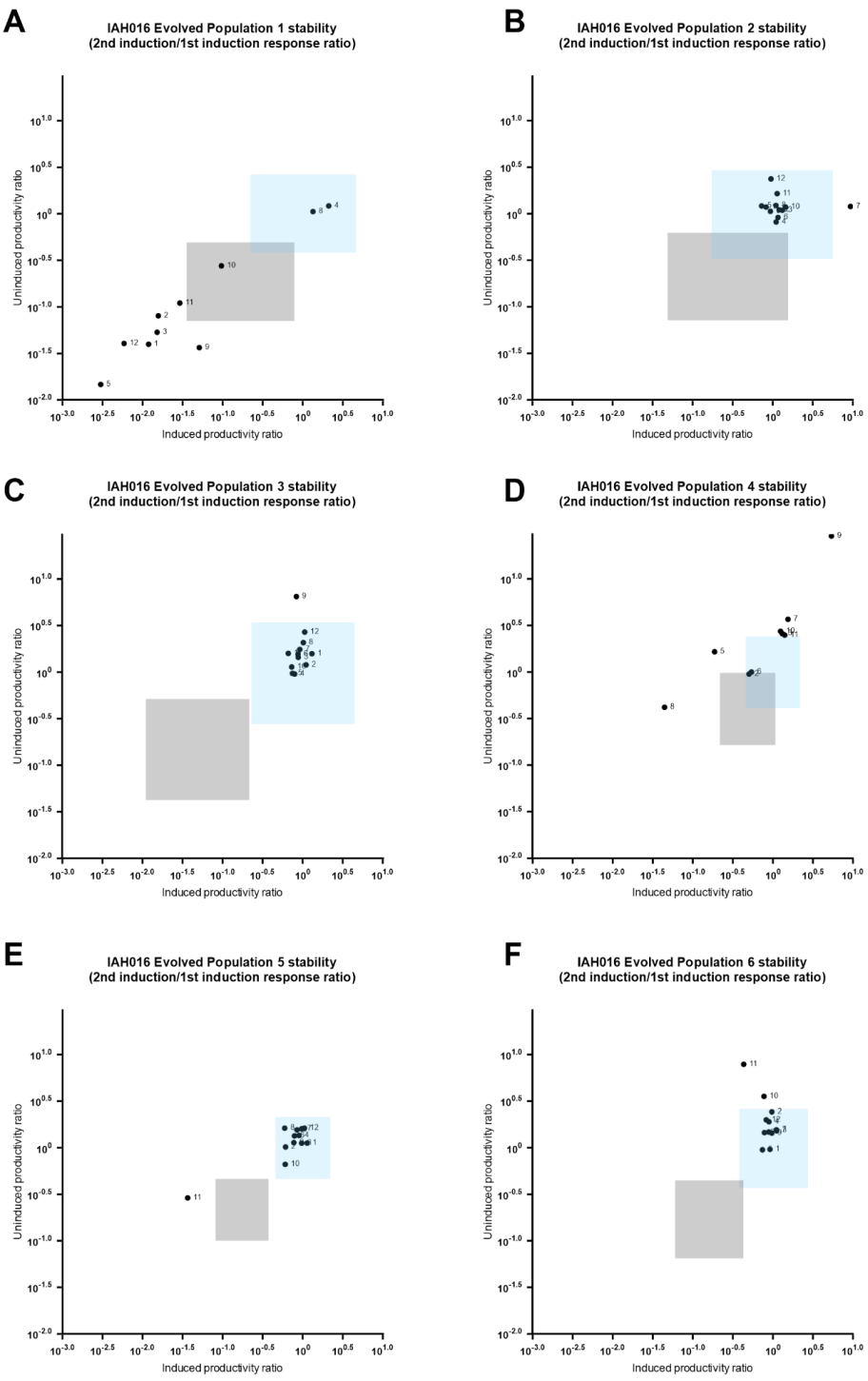


Fig. 4.19 The productivity ratios of evolved IAH016 strains depicting strain stability compared to ancestral means. The grey box shows ancestral stability. The blue box represents a hypothetical, ideally stable population with equal variation as the ancestral population. [9]

[9] The strain designations used in the figures within this chapter do not match the genetic sequencing data obtained later. This is discussed further in Chapter 5.

Most of the clones in four of the six tested IAH016 evolved populations are contained within the stable response (blue) box (**Fig. 4.12B, C, E, F**). This shows that these populations have adapted to the repeated protein overproduction stress. Moreover, three of these populations (**Fig. 4.10C, E, F**) have done so while simultaneously improving either their induced productivity or lowering their uninduced productivity (or both). While population 2 (**Fig. 4.10B and 4.12B**) is likely more stable due to its overall protein output being lower than the ancestors, the genetic adaptation mechanism seen in populations 3, 5 and 6 remains to be determined. Population 1 (**Fig. 4.10A and 4.12A**) has remained phenotypically similar to the ancestral population; it is therefore not surprising that its stability has not improved.

In all 1516 strain populations, the overall stability improved (**Fig. 4.13**). However, it is worth noting that the initial (ancestral) stability frequently overlaps with the hypothetical stable population to a much larger degree than seen in IAH016 strains. This suggests that the BL21 strains' tolerance to plasmid carriage burden differs depending on the plasmid (pAVE011 or pIAH011). Furthermore, unlike the stable populations of IAH016, the clones which maintain productivity during the second induction with IPTG are also shown to express more sfGFP in the absence of the inducer (**Fig. 4.13A and B**). One explanation for this observation is an increased sensitivity of the evolved strains to the inducer concentration in the media, where IPTG used in the first induction experiment carries over in the initial inoculum used in the second induction experiment. This may suggest that the optimal inducer concentration for these populations is below 0.5mM IPTG, which can also be advantageous in the industry setting.

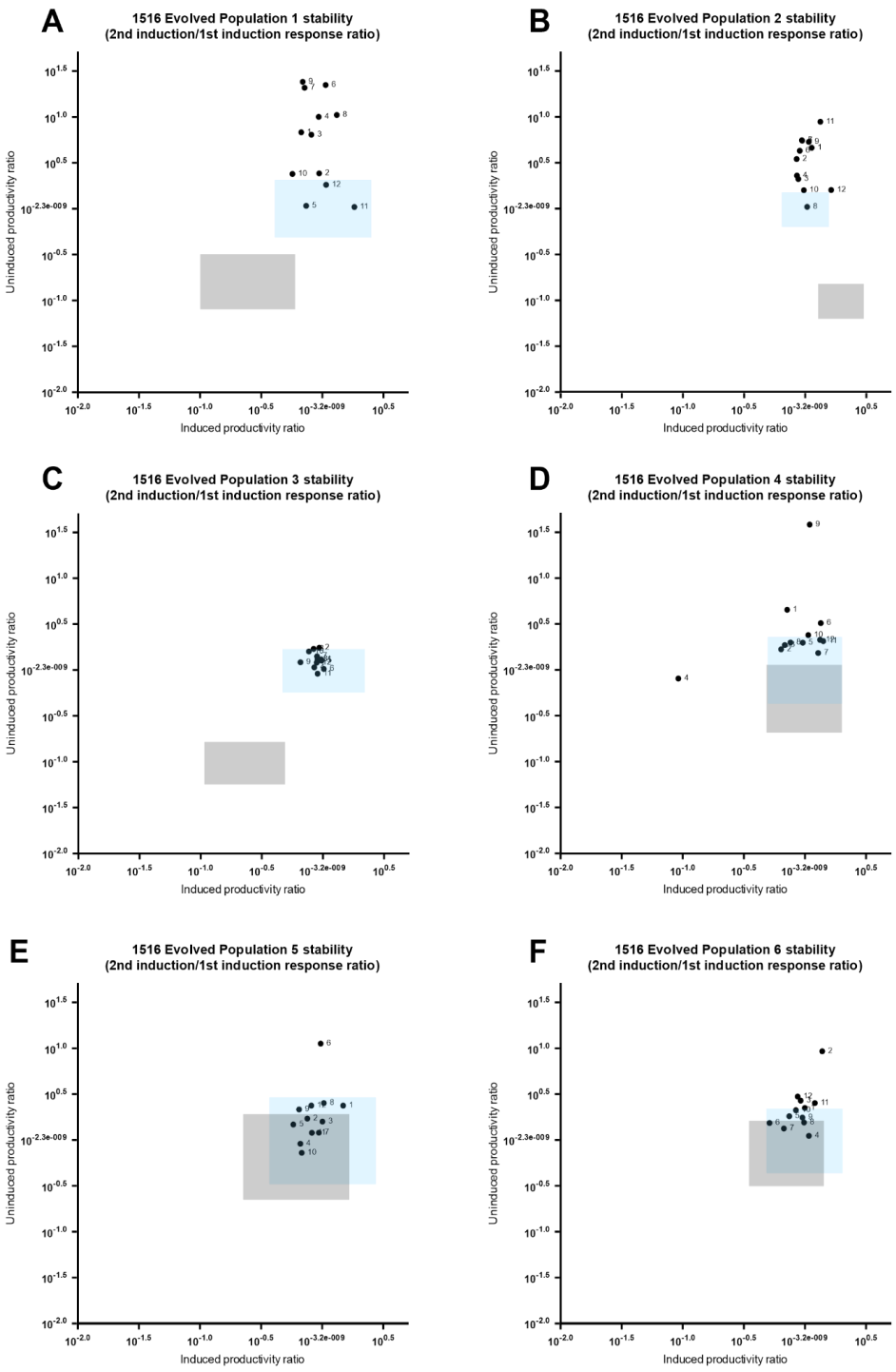


Fig. 4.20 The productivity ratios of evolved 1516 strains depicting strain stability compared to ancestral means. The grey box shows ancestral stability. The blue box represents a hypothetical, ideally stable population with equal variation as the ancestral population. [10]

[10] The strain designations used in the figures within this chapter do not match the genetic sequencing data obtained later. This is discussed further in Chapter 5.

In all IAH015 and 1515 strain populations the stability has improved **(Fig. 4.21)** and similarly to some populations of strain 1516, some clones show increased uninduced productivity during second induction **(Fig. 4.21A, B, C; Fig. 4.22).** The original stability of the strains can be assessed by investigating the overlap between the ancestral and hypothetical, ideally stable populations. The data shows that the ancestral IAH015 populations were more stable than ancestral 1515 populations **(Fig. 4.21 and 4.22).** This trend is the inverse of that observed in B background strains IAH016 and 1516 **(Fig. 4.19 and 4.20).** It suggests that the the burden of plasmid carriage and protein overproduction depends not only on the plasmid, but also on the host genetics: pIAH011 plasmid carriage is better tolerated by the K background *E. coli*, while pAVE011 plasmid carriage is better tolerated by the B background *E.coli*.

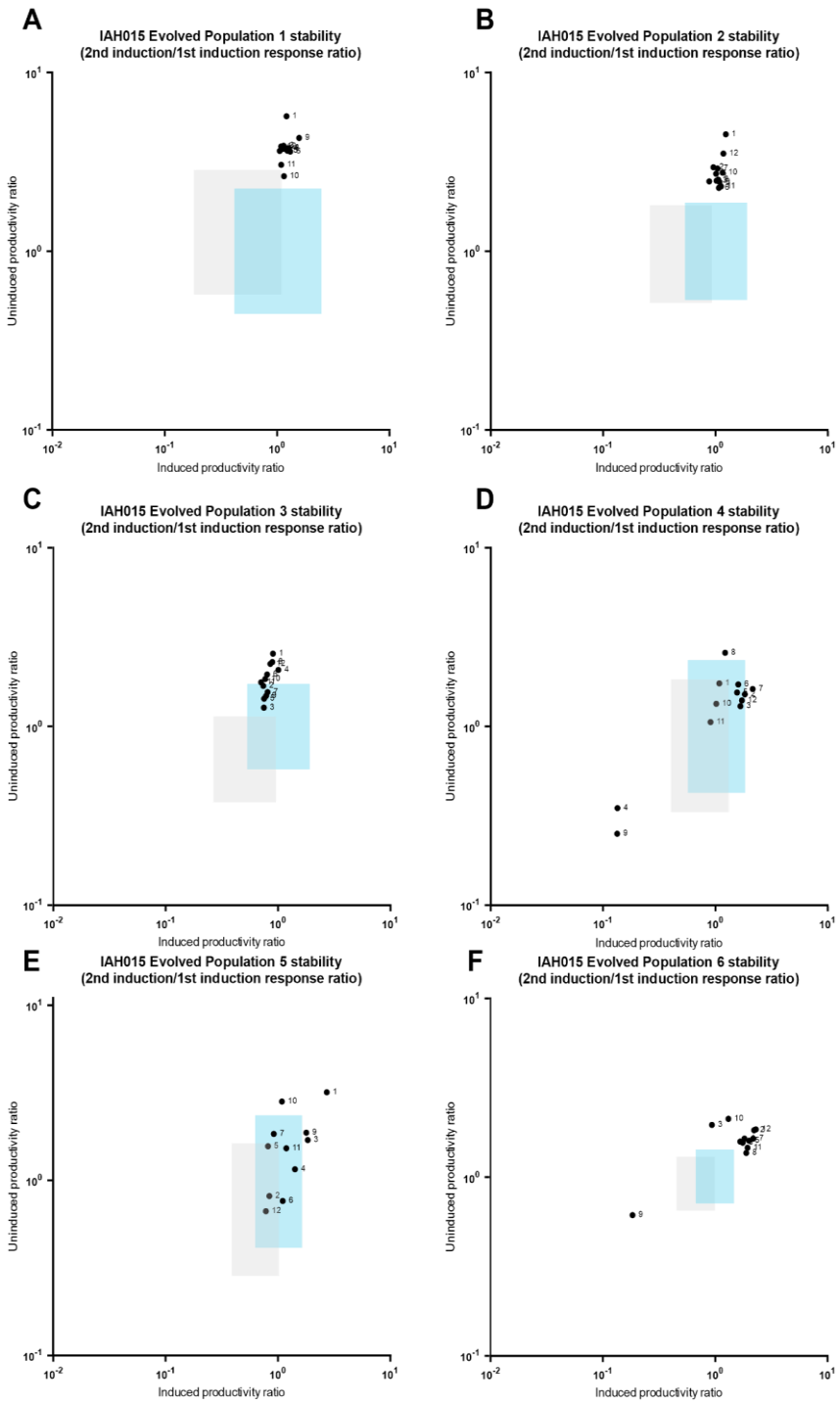


Fig. 4.21 The productivity ratios of evolved IAH015 strains depicting strain stability compared to ancestral means. The grey box shows ancestral stability. The blue box represents a hypothetical, ideally stable population with equal variation as the ancestral population. [11]

[11] The strain designations used in the figures within this chapter do not match the genetic sequencing data obtained later. This is discussed further in Chapter 5.

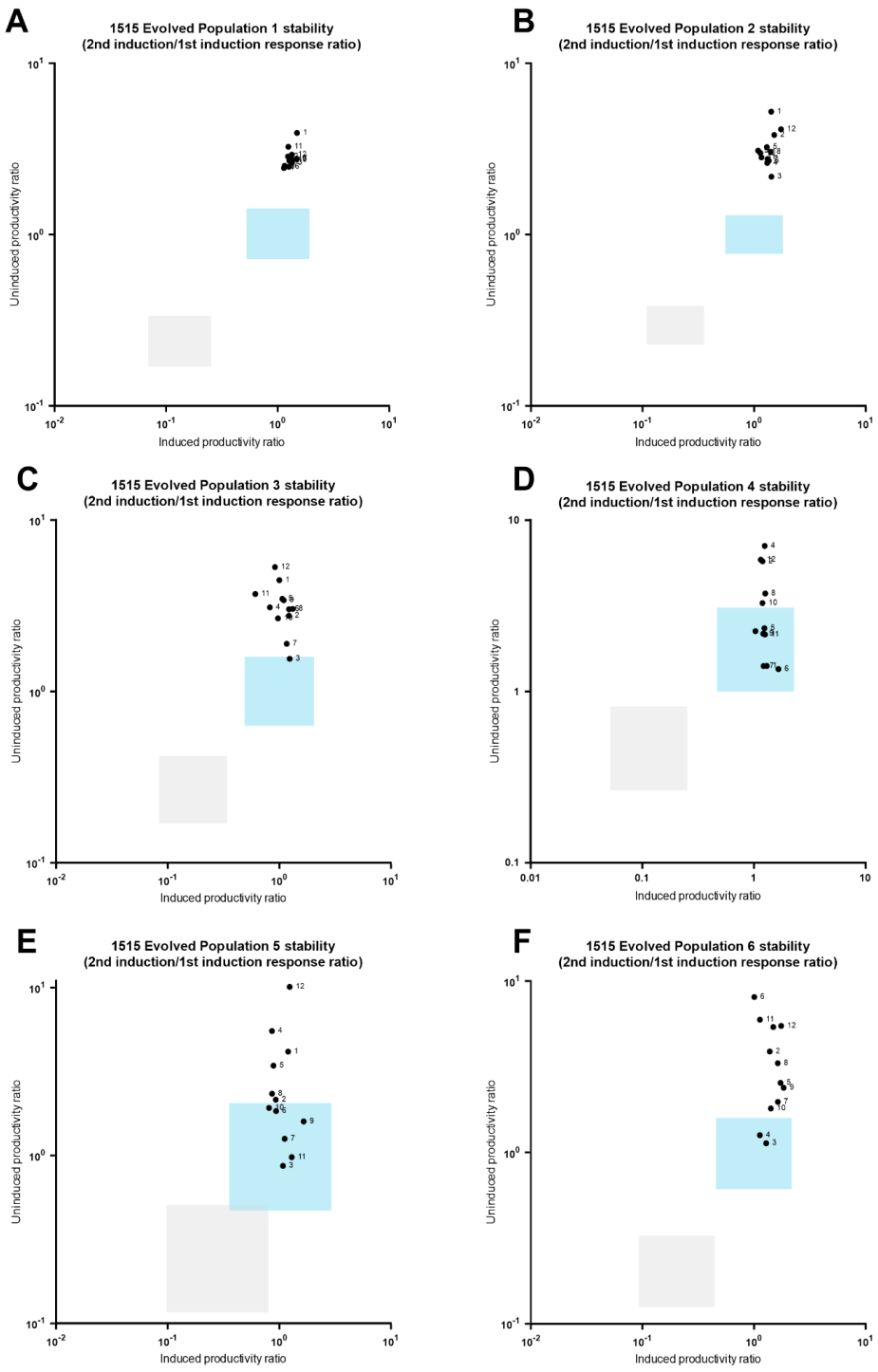


Fig. 4.22 The productivity ratios of evolved 1515 strains depicting strain stability compared to ancestral means. The grey box shows ancestral stability. The blue box represents a hypothetical, ideally stable population with equal variation as the ancestral population. [12]

[12] The strain designations used in the figures within this chapter do not match the genetic sequencing data obtained later. This is discussed further in Chapter 5.

4.2.4 Genomic sequencing clone candidate identification

The next step in identifying the underlying genetic changes responsible for the observed phenotypes was sequencing of both plasmid and the genome. In order for a clone to be identified as suitable for sequencing, two conditions must be met. Firstly, the phenotype must be stable. Using the data on the stability charts **(Fig. 4.19-22)**, any clones with productivity ratio (uninduced or induced) below that of the hypothetical ideally stable population were classified as unstable. Overall, the stability of all four tested strains improved after the evolutionary experiment compared to the ancestors **(Fig. 4.23)** as assessed by the number of stable clones in ancestral and evolved groups.

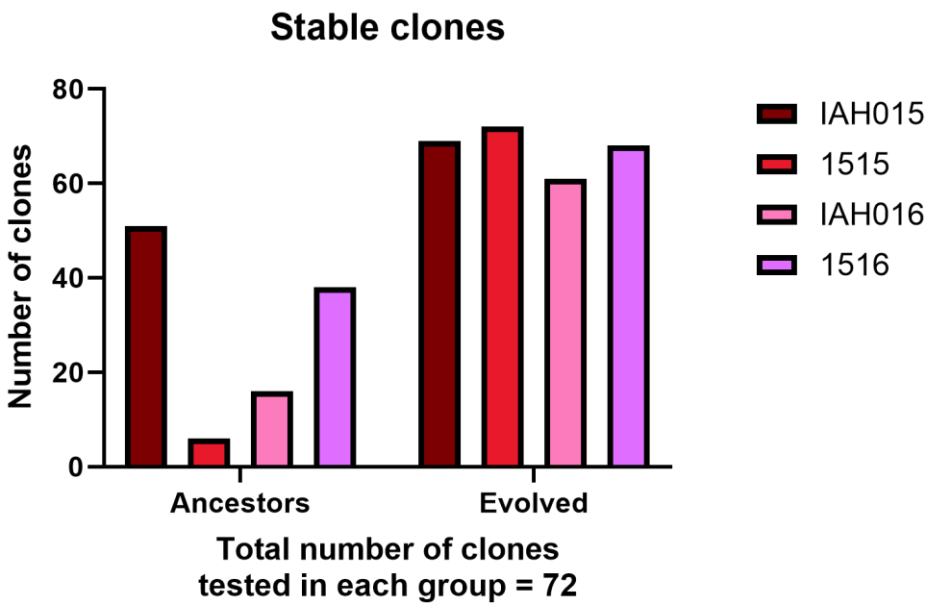


Fig. 4.23 The number of stable clones identified during phenotyping. For each of the four strains, six populations were tested through phenotyping of 12 random clones. The stable clones were identified as clones with 2nd induction/1st induction productivity ratios equal or larger than that calculated for the hypothetical ideally stable population. [13]

[13] The strain designations used in the figures within this chapter do not match the genetic sequencing data obtained later. This is discussed further in Chapter 5.

The second condition required for a clone to be included in the genetic sequencing dataset is their phenotype relative to the ancestors. Five evolutionary groups were identified **(Fig. 4.24A)** based on the evolved phenotypes during the first induction experiments **(Fig. 4.15-18).** Clones in these groups showed an improvement when compared to the ancestors in at least one of the two variables measured (productivity and leakiness).

The distribution of clones across these five evolutionary groups varied depending on the strain genetic background and plasmid combination[14]. For example, all of the "double improved" clones (productivity increased, leakiness decreased; **Fig. 4.24B PULD**) were found in BL21 background, and majority of those in strain 1516, carrying the original pAVE011 plasmid. Meanwhile, the majority of clones in groups representing an evolutionary trade-off (where one variable improves while the other worsens) were found in K background strains IAH015 and 1515 **(Fig. 4.24B PDLD and PULU).** This further supports the hypothesis that the trade-off phenotypes are an intermediate evolutionary step, and that the K background strains are slower to adapt to evolutionary pressures.

Because of the relatively small number of clones present in groups PULU and PULS, all of them were sequenced. Additionally, all five IAH016 clones identified in the PULD group were also sequenced. The remaining groups had a large number of representatives, therefore only some of them were chosen to be sequenced. These clones were evenly distributed between all four strains and all remaining evolutionary

[14] The strain designations used within this chapter do not match the genetic sequencing data obtained later. This is discussed further in Chapter 5.

groups in order to capture as much potential diversity in evolutionary pathways as possible. A total of 96 evolved clones were sequenced.

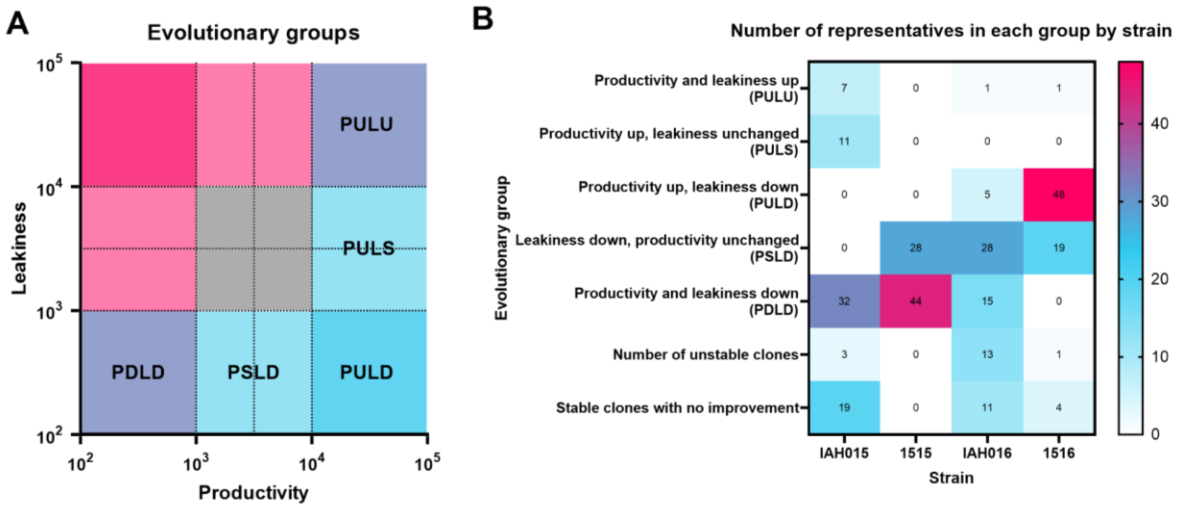


Fig. 4.24 Clones isolated from evolved strains can be categorised into several distinct evolutionary groups. A) The identification of five evolutionary groups showing improvement when compared to ancestors in either productivity or leakiness during the first exposure to the inducer. **B)** The distribution of clones identified in each evolutionary group across the four strains tested.[15]

In order to identify any evolutionary changes, clones from several control groups were also sequenced. From each of the six ancestral populations of strains IAH015, IAH016, 1515, 1516, BL21, W3110 *ΔompT,* BL21 empty pAVE011 and W3110 *ΔompT* empty pAVE011 a single clone was picked at random as a representative to be sequenced. Furthermore, from each of the six evolved populations of strains BL21, W3110 *ΔompT,* BL21 empty pAVE011 and W3110 *ΔompT* empty pAVE011 a single clone was picked at random as a representative to

[15] The strain designations used in the figures within this chapter do not match the genetic sequencing data obtained later. This is discussed further in Chapter 5.

be sequenced. Both ancestral and evolved clones which were sequenced are identified in supplementary tables 6 and 7.

5. Results III "Identifying the genetic variants responsible for improved heterologous protein production profile in pAVEway plasmid-carrying *E. coli* "

5.1 Introduction

In the last two chapters, the plasmid-based *E. coli* heterologous protein production platform pAVEway was characterised and phenotypically unique evolved clones have been identified for further investigation. The clones isolated from evolved *E. coli* cultures carrying either pAVE011 or pIAH011 plasmid were compared to their ancestors via a phenotyping assay. Three phenotypes were identified as relevant in the industrial context. The clones exhibiting higher induced productivity as well as higher leakiness (uninduced productivity) could be used in expression of target proteins non-toxic to the host. The clones exhibiting lower induced productivity as well as lower leakiness (uninduced productivity) might have potential applications in production of difficult to express, toxic to *E. coli* proteins. Finally, the clones showing increased induced productivity as well as reduced leakiness (uninduced productivity) are the most promising with potential applications for protein production at scale.

To understand the genetic changes which contribute to those altered phenotypes, sequencing of both the genome and the plasmid is necessary. Two of the many available sequencing platforms were considered for investigating the genetic changes in the evolved strains - Illumina (via MicrobesNG) and Nanopore sequencing. Generally, a mixture of the two technologies is recommended for the most accurate results (Wick et al. 2023). However, this was not possible in this PhD project due to the number of clones that needed to be investigated and the associated cost.

Illumina sequencing is a short-read based technology. Its main advantage is the high accuracy of the results with an error of only 0.1% (Hu et al. 2021), which was lower than that of long read sequencing technologies until very recently. Its main limitation is the difficulties with accurate de-novo assemblies, particularly involving rearrangements and repetitive sequences, due to relatively short contig lengths (Hu et al. 2021).

Nanopore is a long-read sequencing technology. One of its most significant advantages is its cost-effectiveness, mainly when used directly instead of as a service. However, the long-read sequencing technology has its own drawbacks, mainly higher error rate in homopolymer stretches of the sequenced DNA (Delahaye & Nicolas 2021). However, with recent developments in flow cell chemistry, the errors in base-calling during sequencing has improved in accuracy, reducing the need for additional short-read sequencing, which is often used for polishing nanopore reads (Sereika et al. 2022). An additional advantage of nanopore sequencing is that, unlike the Illumina service provided by MicrobesNG at the time of this study, it can sequence both the genome and plasmid in a single step, provided that the plasmid copy number is high enough in the sample. This can significantly improve the throughput of genetic data acquisition from strains in this study.

Considering the above, a direct comparison approach of both sequencing methods was designed, where four clones (two ancestors and two evolved strains) were sequenced using both technologies. The results were compared to establish whether nanopore sequencing is sensitive enough to detect genotype changes in this study. Following this, the previously identified evolved clones, as well as relevant control groups were sequenced, and their genotypes analysed in context of phenotypes.

5.2 Results and discussion

5.2.1 Direct comparison of short- and long-read sequencing technology accuracy

In order to choose the best sequencing method for this study, two leading NGS technologies were compared. To this end, two ancestral clones isolated from IAH016 (BL21 background strain) strain populations 5 and 6 as well as two evolved clones from the corresponding populations were sequenced using Illumina short-read and nanopore long-read sequencing. The “ancestral” samples were recovered from cryopreserved stocks originating after the first day of the evolutionary experiment. The stocks used to inoculate the plates on the first day were not clonal, but mixed cryopreserved populations, contributing to intrinsic variability within the ancestral populations.

The consensus sequences obtained for each clone were then aligned to a BL21 sequence available online (GenBank: CP010816.1) using MAUVE whole genome aligner plug-in for Geneious Prime. The results suggested multiple (over 40 thousand) single nucleotide polymorphisms (SNPs) even in the ancestral clones. Both W3110 and BL21 strains were used in the same evolutionary experiment, therefore the abundance of SNPs when comparing the consensus sequences to reference BL21 sequence highlighted the possibility of a cross contamination or a sample mislabelling issue.The four consensus sequences were then aligned to a W3110 reference sequence (GenBank: AP009048.1). The resulting alignment showed fewer SNPs in each ancestral clone, and therefore the W3110 was identified as the correct reference sequence.

Because for each of the four clones both long- and short-read sequencing data was available, a consensus assembly for each clone was also created using Polypolish - a tool enabling short-read polishing of the long read sequencing consensus (Wick & Holt 2022).

When comparing the 3 consensus sequences obtained with different methods, some key differences were found compared to online reference. In the ancestral clone IAH016 AR5C7 seven changes were identified. Of those, five were present in all 3 consensus sequences (four substitutions and the expected *ompT* deletion). An additional substitution and a deletion were identified in the nanopore consensus, as well as polypolish consensus, but not in the short-read consensus. In the second ancestral clone, IAH016 AR6C7 a 2 bp deletion in a non-coding region of the DNA was identified only in polypolish and nanopore consensus sequences.

Unsurprisingly, there were more SNPs present in the evolved clones IAH016 ER5C2 and IAH016 ER6C1 than in the ancestral clones. Of 25 changes identified in the ER5C2 clone, 18 were present in all three consensus sequences (short-read assembly, long-read assembly and polypolish assembly using both short- and long-read data). Meanwhile, the remaining 7 changes were only identified in the polypolish and nanopore consensus sequences. In the evolved ER6C1 clone 36 changes were identified. Of those, 15 were identified in all consensus sequences; 20 were identified only in nanopore and polypolish consensus sequences and only one was identified only in the illumina consensus sequence.

These results highlight the high accuracy of sequencing data obtained through long read sequencing. Using short reads for additional polishing of the consensus did not show improvement in identifying changes when compared to a reference sequence, while using short read sequencing results only to create a consensus

resulted in fewer changes being identified. Considering this, nanopore sequencing was chosen as the approach to be used in this study.

5.2.2 Reference genome acquisition

After identifying the clones from both control and evolved groups **(Fig. 4.24)**, the 192 samples were prepared using native barcoding kit 96 (Oxford Nanopore) and sequenced.

The sequencing data for six ancestral samples of W3110 (no plasmid) was assembled *de novo* and this resulted in six circular, full genome contigs. Those 6 genomes were then aligned to each other using MAUVE plug-in for Geneious Prime, and a consensus was extracted from this alignment (identical sites: 99.9%, pairwise identity: 99.98%). The process was repeated for the data obtained from six ancestral BL21 clones (with no plasmid). The BL21 consensus sequence had 99.999% identical sites and 99.999% pairwise identity. The resulting W3110 and BL21 ancestral consensus sequences were then used as reference in variant calling pipeline used to analyse the remainder of the data by the Bioscience Technology Facility at the University of York.

At this stage, it was uncovered that out of 192 samples sequenced, only five were consistent with the BL21 genotype: samples from evolved populations 5 and 6 of plasmid-free BL21 as well as samples from evolved populations 1, 2 and 5 of plasmid-free W3110. When a W3110 (K background) sample data is analysed using a BL21 reference, over 40 thousand polymorphisms are found in each clone. This is inconsistent with the relatively short experimental evolution timeline, as well as the fact that in evolved samples labelled W3110 aligned to a W3110 reference genome no more than two hundred polymorphisms per clone are found. Furthermore, when

plasmid data was analysed, it was also revealed that in all clones of 1516 strain, where the original plasmid pAVE011 sequence was expected, the altered pIAH011 plasmid was present.

The most likely explanation for these results is cross-contamination during the flow cytometry selection. The flow cytometry part of the experiment was always carried out in the same order of samples: 1515, IAH015, 1516, IAH016. There is no evidence for cross-contamination in the 1515 or IAH015 samples. Samples 1516 contained pIAH011 plasmid and were identified as the K background genotype, which was likely introduced into the population through carry over from IAH015 cultures. Samples IAH016 also have K background genotype, likely introduced into the population via carry-over from contaminated samples 1516 sorted just before. This specific direction of cross-contamination between samples suggests that the cell sorting step of selection was the cross-contamination source and clean-up steps between the treatments were not effective enough to remove all bacterial cells from the system. A mislabelling mistake or mistakes in aseptic technique during daily transfers would have likely resulted in cross contamination in all directions, including pIAH011 plasmid in 1515 samples. An additional source of cross-contamination, particularly in the ancestral groups not subjected to the flow cytometry regime could be the cryopreservation step, as it was carried out in 96 well plates instead of individual tubes, with no spaces between different groups.

Even though the sequenced samples labelled 1516 and IAH016 have been found to carry the IAH015 genotype (W3110 *ΔompT* pIAH011), they remained separated during the sequencing data analysis stage. This is because if the cross-contamination hypothesis is true, then those clones would have had to first outcompete the BL21 strains present in the original 1516 and IAH016 cultures. This

competition could have steered the evolutionary adaptations in a different direction than that of 1515 and IAH015 cultures, which evolved only in response to the environment and fluorescent selection. In the following sections of this chapter, those samples will be referred to as 1516* and IAH016*, even though their genotype has been confirmed as the same as strain IAH015. Throughout this chapter, cultures referred to as having "B* genetic background" are actually of K genetic background with possible BL21 evolutionary influences.

5.2.3 Variants identified in gene coding regions of *E. coli*

The variant calling files provided by the Technology Facility at the University of York were imported into the W3110 reference files in Geneious Prime. In order to narrow down the relevant genome and plasmid regions affected by the evolutionary experiment, the names of the affected regions (or closest neighbouring regions for variants found in non-coding regions) were recorded in an Excel spreadsheet, alongside the proportion of clones of specific phenotype in which those variants were found. Additionally, evolutionary groups PULS and PSLD were sequenced, but not analysed. Next, variants representing at least half of the sequenced clones in at least one of the groups were identified and represented in a heatmap format **(Fig. 5.1, 5.4 and 5.5)**. These heatmaps represent only the frequency of the variants and provide no data on the functionality of the genes; however they can be used to identify regions commonly associated with specific phenotypes. These commonly affected regions were identified as represented in at least 0.7 of the clones in a given evolutionary group, as outlined in Fig.4.24.

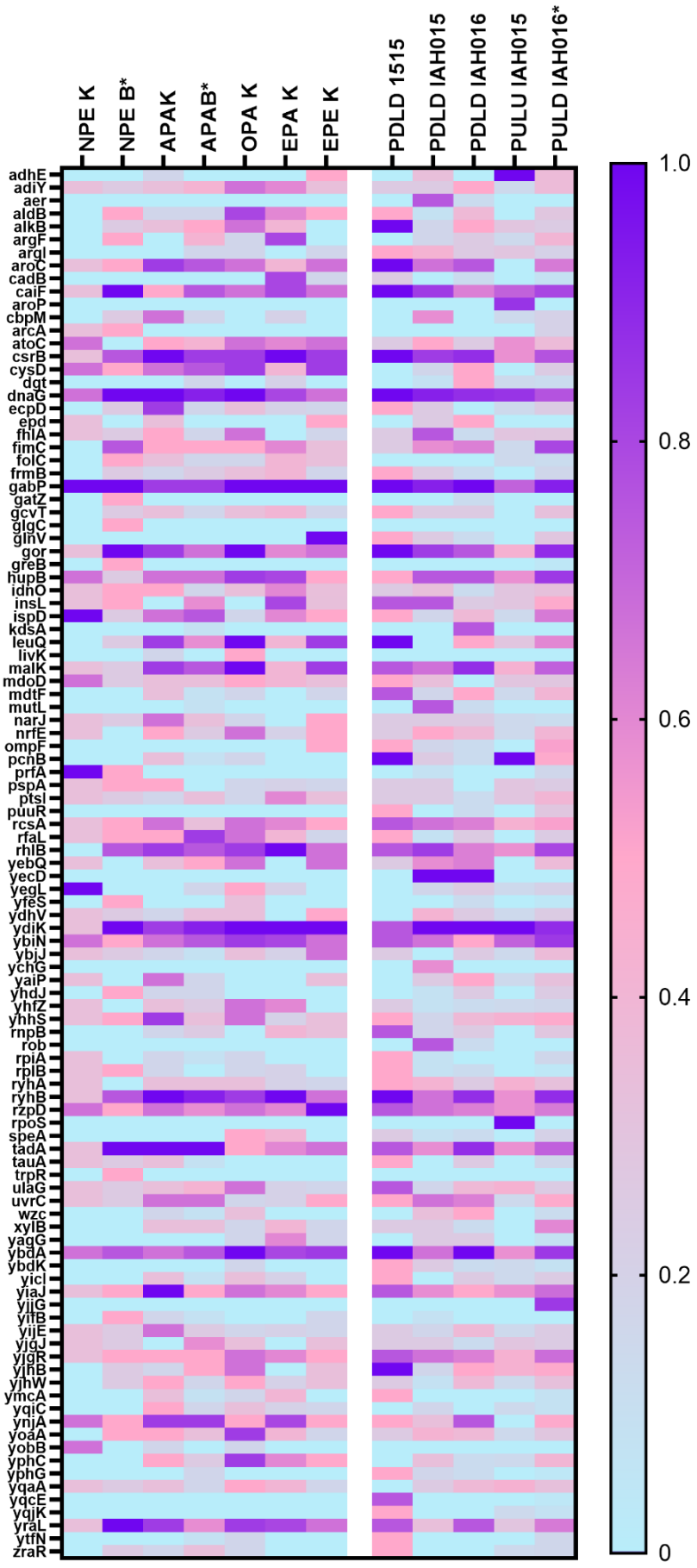


Fig. 5.1 The frequency of genetic changes affecting specific protein coding regions of the *E. coli* genome in several phenotypic groups when compared to no plasmid ancestor reference genome. Strains labelled with a * have been labelled as BL21 throughout the experiment, but have been now identified as W3110 genotype. The scale represents the fraction of each group where a polymorphism was identified in a given region; 0=present in no clones in a group, 1=present in all clones in a group. Group sizes were varied: no plasmid evolved K background (NPEK) = 3 clones; no plasmid evolved B* background (NPEB*) = 4 clones; altered plasmid ancestral K background (APAK) = 6 clones;

altered plasmid ancestral B* background (APAB*) = 12 clones; original plasmid K background (OPAK) = 6 clones; empty plasmid ancestral K background (EPAK) = 6 clones; empty plasmid evolved K background (EPEK) = 6 clones; productivity down leakiness down (PDLD) 1515 = 4 clones; PDLD IAH015 = 12 clones; PDLD IAH016* = 8 clones; productivity up leakiness up (PULU) IAH015 = 7 clones; productivity up leakiness down IAH016 = 25 clones.

5.2.3.1 Protein coding regions affected in rapid (2-5 passages) adaptation to media

Even though no polymorphisms were expected in the control ancestral groups (EPAK, APAK, APAB*, OPAK) when compared to a no plasmid ancestor W3110 *ΔompT* consensus sequence, several regions show a clear high proportion of clones affected **(Fig. 5.1).** Given the number of passages to fresh media and growth required for transformation with a plasmid or propagation is between 2 and 5, those changes detected in the control groups must be occurring rapidly and are most likely in response to media components. This hypothesis is supported by the differences in media components used by FujiFilm and the vLB recipe used during the evolutionary experiment.

Twenty-nine regions were identified as affected by those media adaptation changes **(Table 5.1).** Of those four (*ynjA, yphC, ydiK, yhhS*) are poorly characterised and the impact of the polymorphisms is difficult to predict. Another region with unknown impact is the transposase *insL*. The rest of the impacted genes can be divided into six distinct functional groups, which will be discussed in the following sections of this chapter.

Gene name	Nature of the polymorphism (most common)	Potential effect on protein functionality
aldB	K366del	unknown
argF	K54X	truncation
aroC	R198_L200delinsAK	unknown
cadB	K5fsX19	Frameshift, premature stop codon
caiF	A27fsX44	Frameshift, premature stop codon
csrB	Multiple T deletion	Unknown; affects 3' end of RNA
cysD	K161_N162delinsK	unknown
dnaG	P162del	unknown
yadV(ecpD)	K41fsX246	unknown
gabP	H118fsX141	frameshift
gor	K430del	Change within dimerisation domain of the protein
hupB	L44fs53	Frameshift, premature stop codon
insL	Variable SNP (start amino acid 359)	unknown
ispD	G131fsX141	frameshift
leuQ	various	unknown
malK	A105fsX120	frameshift
rfaL	K131fsX144	Frameshift, premature stop codon
rhlB	L71fsX94	Frameshift, premature stop codon
ryhB	A deletion	Unknown; affects 3' end of RNA
tadA	K172fsX219	frameshift

entS (ybdA)	F262fsX275	frameshift
rlmF (ybiN)	K277del	unknown
ydiK	M133_G134delinsM	unknown
yhhS	L390_K391delinsL	unknown
yiaJ (plaR)	G7fsX33 and G7_K8delinsG	Frameshift and premature stop codon in 74% of clones showing the mutation
ynjA	W34_G35delinsW	unknown
yoaA	K63_K64delinsK	unknown
yphC	2 variants L143_L144delinsL or I142fsX164	unknown
yraL (rsmI)	K260fsX271	Frameshift, premature stop codon

Table 5.1. The list of coding regions frequently affected by mutations in the control groups of the experiment. The table shows the mutation position, as well as the amino acid changes and potential effect on the protein functionality.

5.2.3.1.1 Small regulatory RNAs

Two regions (*csrB* and *ryhB*) encode small regulatory RNAs. CsrB binds to CsrA protein and inhibits its functionality as a regulator of the carbon metabolism (Liu et al. 1997), while *ryhB* acts to limit iron consumption in iron-poor environments (Massé et al. 2005) The detected polymorphisms were found in regions encoding the 3' ends of both RNAs, which contain homopolymer stretches of unpaired uracil residues. This is important for two reasons. Firstly, nanopore sequencing has a disadvantage of higher rate of mistakes in homopolymer stretches of DNA when compared to short-read sequencing. Therefore, the sequence of both *csrB* and *ryhB* in the clones of the control groups should be confirmed by short-read sequencing before considering the implications of those changes. Secondly, the uracil 3' ends of both RNAs do not take

part in their secondary structure formation, which is relevant for their interaction with target proteins (Liu et al. 1997)(Peterman et al. 2014). It is therefore unlikely that those polymorphisms have any effect on the functionality of the *csrB* and *ryhB.*

5.2.3.1.2 Metabolism genes

Six of the frequently affected regions in response to media adaptation related to *E. coli* metabolism. Of those, polymorphisms in aldehyde dehydrogenase B (AldB), chorismate synthase (AroC) and sulphate adenylyltransferase subunit 2 (CysD) affect protein regions not implicated in functionality. Furthermore, they do not cause frameshifts or premature truncation, and therefore their impact on the phenotype is difficult to predict.

The mutations detected in glutathione reductase (*gor*) do not cause a frameshift or introduce a premature stop codon. The change is a 3 base pair deletion removing Lysine at the 430 amino acid position. Glutathione reductase is a homodimeric enzyme, and amino acids at positions 339-449 are indicated in its dimerisation (Finn et al. 2016; Mittl & Schulz 1994). It was previously reported that changes to amino acids in the dimerisation domain of the protein may affect the dimer stability and enzyme kinetics (Bashir et al. 1995). However, it was also previously reported that glutathione reductase is not essential for maintenance of the reduced glutathione pool in *E. coli* (Tuggle & Fuchs 1985). This suggests that even if the mutation observed in this study adversely affected the glutathione reductase enzyme, it would be unlikely to impact the viability of the strain. It is also possible that an altered kinetic profile of the enzyme is more beneficial for growth in vegetable LB media.

The second gene involved in *E. coli* metabolism which was found to be altered in response to media adaptation is ornithine carbamoyltransferase (*argF)* which

catalyses the arginine biosynthesis (Legrain et al. 1976). The observed mutation results in a truncated protein that consists of 54 residues. . The resulting protein lacks crucial arginine 57 residue, involved in substrate binding and catalysis (Kuo et al. 1988). Therefore, the mutation in *argF* observed as response to the media components is almost certainly a loss of function mutation. Losing functionality of one of the biosynthetic pathways may be beneficial for bacteria grown in rich media, as it allows for redirecting resources from a redundant, energy consuming pathway towards other processes.

The last metabolism related gene affected in the control groups of this experiment is *ispD,* encoding the 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase. Numerous studies have shown that this gene is essential for *E. coli* growth (Campos et al. 2001; Freiberg et al. 2001; Baba et al. 2006). A frameshift mutation identified in the control groups of this study would likely result in a non-functional protein. The identified mutation is a single G deletion in a 7 bp long homopolymer stretch and may therefore be an artefact related to nanopore sequencing rather than a genuine polymorphism.

5.2.3.1.3 Transporter genes

Four transporter genes have been identified as affected by mutations resulting from rapid adaptation to the media. All of the identified mutations have been determined to cause a frameshift resulting in a premature stop codon, and therefore are hypothesised to result in complete loss of function.

The first of the affected genes, *cadB,* encodes a lysine:cadaverine antiporter. Its function is important for acid resistance (Kanjee & Houry 2013) and under acidic conditions it facilitates lysine uptake from the environment and cadaverine excretion

into the environment. However, the expression of the *cadB* under neutral pH conditions has been shown to be negligible (Soksawatmaekhin et al. 2004).

The second of the transporter genes, *gabP*, encodes a 4-aminobutanoate:H+ symporter (Niegemann et al. 1993). The mutation in this gene suggests that when grown in vegetable LB, *E. coli* is not relying on γ-aminobutyric acid as a source of carbon and nitrogen, as it is likely using the nitrogen in regular amino acids present in the media.

The next transporter-related gene showing a frameshift mutation is *malK,* one of the maltose transporter subunits. Maltose is not the preferred source of carbon in *E. coli*, which follows a sugar utilisation hierarchy (Aidelberg et al. 2014). This hierarchical regulation of sugar uptake and utilisation genes allows for energy and resource conservation by expressing only the genes required to process the carbon source currently used by the cell.

Finally, the last of the affected transporter encoding genes, *entS* (also known as *ybdA*) is a enterobactin exporter implicated in arabinose efflux (Koita & Rao 2012) and alleviating iron deprivation (Furrer et al. 2002). Iron deprivation should not be an issue in these media conditions.

Considering the functions of the transporter genes affected by frameshift mutations resulting in premature stop codons, it is likely that these mutations allow for conservation of cellular resources by interrupting synthesis of proteins which are redundant in rich media (such as vegetable LB).

5.2.3.1.4 DNA binding proteins and regulators

There were five regions commonly affected in response to media adaptation which encoded DNA binding proteins and transcription factors.

The first of the affected genes encodes DNA primase *(dnaG*) and has been reported as an essential gene (Baba et al. 2006; Gerdes et al. 2003; Goodall et al. 2018). It is responsible for RNA primer synthesis on single stranded DNA (Rowen & Kornberg 1978). The observed mutation is a 3 base pair deletion, removing glycine at position 162. This position has not been reported as essential for the primase functionality, and therefore the effects of the mutation are difficult to predict. However, based on the fact that the gene is essential for *E. coli* growth this mutation is unlikely to disrupt the protein function.

The second of the affected genes, *yoaA*, encodes an ATP-dependent DNA helicase implicated in DNA damage repair and protection against certain antimicrobial agents (Brown et al. 2015). The mutation occurs in the ATP-binding region of the protein and causes a deletion mutation. The original sequence 5'-AAAAAGAAA-3' becomes 5'-AAGAAA-3', and this changes the amino acids from 3 lysines to 2 lysines. Because of this, the rest of the amino acids in the protein are not affected and its functionality is most likely preserved. However, impaired function of a gene involved in DNA repair could also be beneficial for the *E.coli* adapting to new environments. DNA damage repair systems have been shown to affect the rates of mutagenesis in *E. coli,* and mutations are a driver of evolutionary adaptation (Foster et al. 2015).

The next three proteins are transcription regulation factors. The first factor, CaiF*, is very* specific in its DNA binding (Buchet et al. 1999) and recognises only one site. It activates the genes involved in carnitine metabolism under anaerobic conditions (Eichler et al. 1996). During the evolutionary experiment, oxygen was available. Therefore, the carnitine metabolism genes would not be transcribed. A mutation in the transcription regulator *CaiF* would ensure that those redundant genes remain untranscribed and the cellular resources can be diverted to other processes.

Another transcription regulator found to be mutated in the control groups of the experiment is YiaJ (also known as PlaR). YiaJ is a repressor negatively controlling the expression of *E. coli* genes metabolising rare, plant derived nutrients (Shimada et al. 2019). It has been reported that disrupting *yiaJ* results in constitutive expression of genes regulated by it (Badía et al. 1998). In the evolutionary experiment, vegetable LB was used, which contains vegetable tryptone. While the exact composition of that media component is not specified by the manufacturer, it contains plant derived nutrients. Since *yiaJ* mutants show better growth on plant-derived carbon sources than mutants expressing *yiaJ* (Shimada et al. 2019), this mutation is likely the cause for the faster growth of the evolved strains in this study in response to growth in vLB media.

The final transcription regulator gene *hupB* mutated in the control groups of the evolutionary experiment encodes a DNA binding protein HU-β. It is a subunit of the DNA-binding transcriptional dual regulator HU which represses expression of *pgm (*phosphoglucomutase) and *gal* genes, both of which are involved in galactose utilisation. A disruption in *hupB* functionality would likely upregulate those genes and allow for their expression regardless of the sugar utilisation hierarchy. This may provide growth advantage depending on the media composition. The *hupB* encoded protein also has more general functions as a modulator of DNA packaging.

5.2.3.1.5 Cell envelope composition genes

Only two of the genes involved in the cell membrane composition were identified as mutation hotspots in response to media adaptation. One of them (*ecpD*, also known as *yadV)* is a putative fimbrial chaperone. While the mutation occurs in the N-terminal domain of this protein, it is difficult to predict its impact on protein functionality. However, fimbrial proteins are known to be costly to produce and not

beneficial outside of specific environments; furthermore, deletion of the fimbrial genes improves other protein production in *E.coli* (Qiao et al. 2021). Therefore, if the mutations in *yadV* are disrupting its function, it would provide the benefit of freeing up the cellular resources necessary for growth in the ancestral control groups of the evolutionary experiment.

Another gene affecting the composition of cell membrane has been found to be affected by mutations in the ancestral control groups - *rfaL.* This gene encodes O-antigen ligase (Whitfield et al. 1997). The mutation discovered in the ancestral groups of this study causes a frameshift and a premature stop codon. However, in K background *E. coli* strains, disruption of *rfaL* function does not produce a distinct phenotype (Roncero & Casadaban 1992), since K background *E. coli* does not produce O-antigen due to a disruption in the *rfb* gene cluster. In this study, the disruption of this redundant gene may contribute to better resource management of the *E. coli* growing in vLB media.

5.2.3.1.6 RNA and RNA-associated genes

The last group of genes affected in the ancestral control groups of the evolutionary experiment are RNA and RNA-associated genes.

The first of the affected genes, *leuQ*, encodes one of the eight leucine tRNAs. The polymorphism nature is either a single C deletion or insertion in a homopolymer stretch of 8 C nucleotides in the 3' end of the coding sequence. This mutation is therefore most likely an artefact caused by the nanopore technology used to sequence the samples. Similarly, a single A deletion was detected at high frequency in the *tadA* gene, in a homopolymer stretch containing 6 A nucleotides. This gene encodes a tRNA adenosine34 deaminase and has been shown to be essential for *E. coli* growth

(Gerdes et al. 2003; Baba et al. 2006; Goodall et al. 2018). Therefore, it is also likely that this polymorphism is caused by the inaccuracies in base-pair calling by the nanopore technology in homopolymer DNA stretches.

ATP-dependent RNA helicase RhlB is the next target for a frequent frameshift mutation resulting in a premature stop codon in the control group samples. This protein is a crucial component of the degradosome, responsible for unwinding of double stranded RNA molecules (Py et al. 1996; Coburn et al. 1999). A mutation in this gene causes no growth defect in rich medium; furthermore, in vitro assays have shown that the RhlE can functionally replace RhlB (Khemici et al. 2004; Jagessar and Jain 2010). Therefore, a mutation acquired in the ancestral groups of this experiment would benefit *E. coli* by preventing translation of a potentially redundant in the vLB environment protein.

Finally, two RNA methyltransferase genes have been identified as targets for mutation in adaptation to growth in vLB media - *ybiN,* also known as *rlmF*, as well as *yraL,* also known as *rsmI*. RlmF is the methyltransferase responsible for methylation of 23S rRNA at the N6 position of the A1618 nucleotide, and mutations have been shown to cause a growth defect in previous studies (Sergiev et al. 2008), although this defect in rich media without an additional heterologous protein expression burden was negligible (Pletnev et al. 2020). Furthermore, the mutation found in this study does not cause a frameshift or introduce a premature stop codon, therefore the impact of the mutation is difficult to predict. RsmI is the 16S rRNA 2'-O-ribose C1402 methyltransferase, and the reported effect on growth of *rsmI* are mixed and range from no defect to slight defect (Kimura and Suzuki 2010; Arigoni et al. 1998; Dassain et al. 1999). While the mutation found in the control ancestral groups of this study does cause a premature stop codon, it only affects the last 28 amino acids of the protein.

5.2.3.2 Protein coding regions affected in slow (7 weeks) adaptation to media

The next group of mutations was identified in the NPE K and NPE B* groups (no plasmid evolved). In these groups, several of the mutations described in the previous section were present, alongside three additional regions. Therefore, those three regions were identified as involved in the adaptation to growth in vLB media long-term. Of those three genes, one (*yegL*) is poorly characterised and no hypothesis can be formulated in regards to its impact on the evolved phenotype.

The mutation in *fimC* gene encoding a type 1 fimbriae periplasmic chaperone causes a frameshift and a premature stop codon. Much like the previously described mutation in fimbrial chaperone *yadV,* the disruption of this gene can be a beneficial adaptation which redirect the cellular resources from fimbriae biogenesis to other processes.

The mutations detected in the *prfA* gene were single nucleotide substitutions, mostly resulting in single amino acid changes (Leucine 340 to Proline, Glutamic acid 178 to glycine). The *prfA* encodes peptide chain release factor RF1 and has been identified as essential for growth of K background *E. coli* (Johnson et al. 2012). None of the mutations detected affect any of the amino acids crucial for protein functionality, therefore these mutations are unlikely to cause any phenotype changes or provide a growth advantage to the plasmid-free cells evolved in vLB media.

5.2.3.3 Protein coding regions affected by adaptation to empty plasmid carriage

In the next analysed group, EPE K, there were several mutations detected which were consistent with adaptations to growing in vLB media. However, there were also two regions affected in this group which were likely involved in the adaptation to

empty plasmid carriage. The first one was a synonymous single nucleotide polymorphism in the *glnV* gene encoding one of the glutamine t-RNAs. This mutation is likely not relevant for the evolved phenotype.

The second region affected by mutations is the *rzpD* gene encoding a putative prophage endopeptidase RzpD. Out of 6 affected clones, the mutation in 3 of them causes a frameshift mutation. The disruption of the DLP12 prophage cassette has been shown to lead to an increased production of outer membrane vesicles (OMVs) in *E. coli* (Pasqua et al. 2021). OMV production is usually increased as a response to the cell envelope stressors (Pasqua et al. 2021). It is possible that empty plasmid carrying *E. coli* cells are adapting to the metabolic burden of plasmid carriage by increasing the OMV production through *rzpD* disruption.

5.2.3.4 Gene variants identified in phenotypically diverse evolved E. coli groups

After identifying the variants in gene coding regions responsible for adaptations to growth in vLB as well as empty plasmid carriage, the remaining genes affected in each of the evolved groups were analysed.

5.2.3.4.1 Productivity up, leakiness up phenotype (PULU)

The first phenotype was represented by clones showing increased sfGFP production in presence as well as absence of the IPTG inducer. In those clones, two protein coding regions were found to be affected by mutations in addition to previously identified media and plasmid carriage adaptation variants. In the first region, *adhE,* the mutation was a single nucleotide polymorphism, a synonymous mutation. Since a synonymous mutation cannot change the protein functionality, this mutation is likely irrelevant in the PULU phenotype.

The second of the affected genes is *pcnB* encoding poly(A) polymerase I. This protein is crucial in plasmid copy number maintenance of ColE1 derived plasmids, such as pAVE011 and pIAH011. Poly(A) polymerase I promotes the decay of RNA I, an unstable antisense RNA which is complementary to the RNA II primer required for ColE1 plasmid replication (He et al. 1993). Because of this mechanism, a functional *pcnB* is required for ColE1 plasmid maintenance, and its disruption causes reduction in plasmid copy number (Lopilato et al. 1986). The mutation uncovered in the clones of PULU phenotype is a 7bp deletion causing frameshift starting at 313 amino acid residue of the protein. If this affected the protein function negatively, the expected phenotype of the clones would involve reduced productivity; however this is not the case. Instead, the productivity of those clones is increased. In previous studies, it has been reported that the C-terminus of PcnB is not required for protein functionality (Liu & Parkinson 1989). Here, the *pcnB* mutation in combination with increased productivity suggests that this frameshifted, truncated protein variant may increase the plasmid copy number of the pAVE011 and pIAH011 plasmids. The catalytic component of the protein is in the neck domain located in the N-terminus and the body region of the protein form an RNA binding surface in the protein **(Fig. 5.2).** This region differs in different bacterial and eukaryotic poly A polymerases suggesting that the observed frameshift would likely modulate activity, hypothetically upwards, rather than decreasing it through loss of function (Toh et al., 2011) . However, this hypothesis would need to be tested experimentally. This could involve estimating the plasmid copy number in the ancestral strains as well as the isolated PULU evolved clones. Another way to test this hypothesis would be in vitro testing of the mutant protein functions.

A recent study identifying mutations that increased butanol production in *E. coli*, identified a pcnB (R149L) mutation that increased butanol production. The authors did not consider the role of this mutation in plasmid copy number, even though their expression strain contains a ColE1 origin-based plasmid, pBBR1-aceEF.lpd, rather assuming that the increased productivity was due to change in the stability of other mRNAs in the cell, which is also a possible alternative mechanism (Davis et al., 2023). Also, a very recent paper from Francis et al (2023) has demonstrated that loss of PcnB can lead to general increased in stress tolerance to *E. coli* production strains, which is consistent with the Davis data. This study also suggested that this response is independent on RpoS adaptations to stress. Indeed, it could be that the loss of *pcnB* function is operating in the PULU strain to increase stress tolerance (even if no 'gain of PcnB function is occurring).

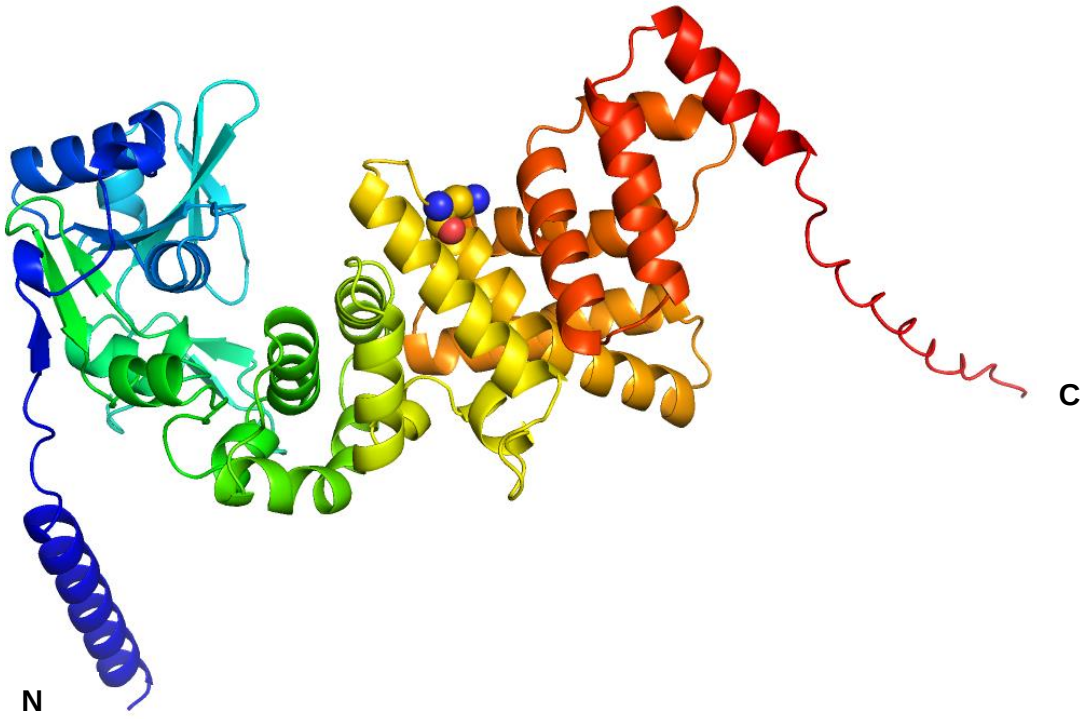


Fig. 5.2 Predicted structure of *E. coli* PcnB, coloured from the N-terminus (blue) to the C-terminus (red), showing the position of the putative 'gain of activity' frameshift at Asn313 (show in space filling). Prediction created with AlphaFold.

5.2.3.4.2 Productivity down, leakiness down phenotype (PDLD)

The next phenotype was represented by clones showing decreased sfGFP production in presence as well as absence of the IPTG inducer.

The most striking of the fifteen regions frequently affected in this group is *pcnB*. The mutations found in this group are heterogenous - some are insertions, some are substitutions. However, in contrast to the mutations found in *pcnB* in the group with opposite phenotype (PULU), those found in the PDLD group are all affecting the N-terminal amino acids (specifically within the first 250 amino acid residues). As discussed previously, it is the N-terminus of the protein which is crucial to its function, therefore the mutations in PDLD group are likely to interfere with *pcnB* product function. This would in turn reduce plasmid copy number of the pAVE011 and pIAH011 plasmids, resulting in productivity and leakiness to decrease. Mutations in this region of *pcnB* have been previously identified as loss of function mutations (Yuan et al. 2024).

Three out of the fifteen regions affected in the PDLD group are poorly characterised (*yecD, yjgR* and *yqcE*), therefore no predictions about their role in the PDLD phenotype can be made at this time. Another four genes which did not fit into any of the categories discussed below were investigated (*ulaG*, *aer*, *rnpB* and *kdsA*), however the mutations uncovered were unlikely to contribute to the PDLD phenotype. The mutations in *aer* and *kdsA* genes were single nucleotide polymorphisms, altering one amino acid residue outside of any identified regions crucial for protein functionality. Meanwhile, the mutations in *rnpB* and *ulaG* involved a single or double nucleotide deletion in homopolymer stretches of the coding sequences, therefore the confidence in those mutations is low given the limitations of nanopore sequencing technology.

5.2.3.4.2.1 DNA damage repair genes

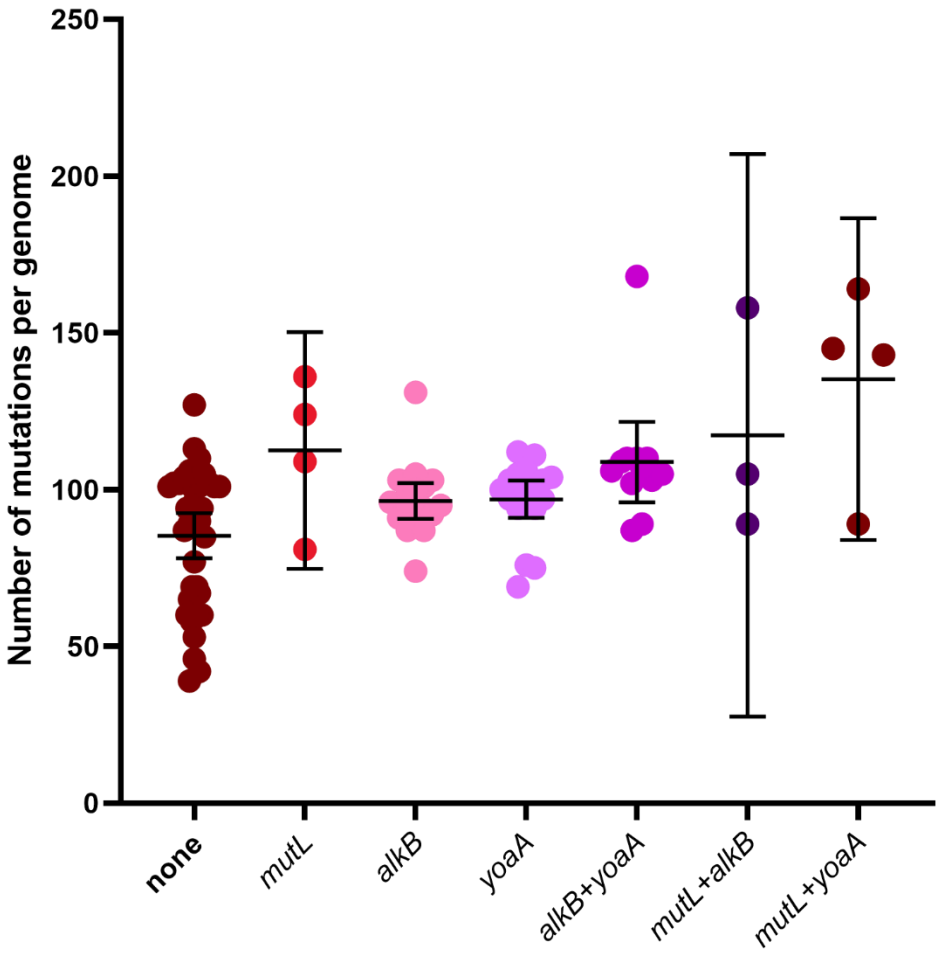
Two of the affected regions encode DNA damage repair genes: *alkB* and *mutL.* The mutation found in *alkB* is a 3bp deletion, resulting in Lysine 215 residue being removed from the encoded protein DNA oxidative demethylase. The demethylase normally repairs methyl lesions in the DNA and RNA, and the affected amino acid residue is not crucial for its function. The second of the affected genes, *mutL,* contains a 6bp deletion (5'-GCTGGC-3') resulting in the deletion of Lysine 68 and Alanine 69 residue deletion. MutL is a mismatch repair protein with a highly conserved N-terminal domain (Kosinski et al. 2005). While the amino acid residues discovered to be missing in the MutL*L* variant present in PDLD clones are not directly within highly conserved motifs, they are located between two such motifs **(Fig. 5.3)** (Banasik & Sachadyn 2014). The motif directly before the mutated residues interacts directly with ATP and is crucial for the protein's ATPase activity (Ban & Yang 1998; Banasik & Sachadyn 2014). The motif located after is one of the three disordered loop motifs, involved in protein conformation switching allowing ATP binding (Ban & Yang 1998; Banasik & Sachadyn 2014). The observed mutation might therefore interfere with the MutL ability to bind ATP. ATPase activity is required for the MutL protein to perform its function in DNA mismatch repair (Robertson et al. 2006). The same mutation has been previously identified in a REL606 (B background) strain of *E. coli* and linked to protein loss of function and a mutator phenotype (Shaver et al. 2002). This in-frame mutation was also found to provide a fitness advantage when competing with strains with functional MutL, leading to the mutant allele fixation in the population despite no evidence for direct advantage caused by resource conservation. Since a mutator phenotype is characterised by an increased number of random mutations, it can be beneficial in adapting to stressful conditions. Indeed, this has been reported in the literature,

specifically highlighting the fitness advantage of clones with medium-high mutation rates compared to their ancestors (Sprouffske et al. 2018).

Fig. 5.3 A schematic representation of *E. coli* W3110 MutL sequence. The amino acid residues are represented with the letters and numbered on the top of the figure. The secondary structure elements are represented with the yellow arrow (beta strand) and magenta bar (alpha helix). The conserved motifs are underlined. The emboldened letters represent the amino acids removed by a 6bp deletion in the coding sequence of the gene in the PDLD phenotype experimental group. The figure was created with Biorender.com.

In addition to *mutL* variant identified in the PDLD group being responsible for the strain's increased adaptability, it is likely also responsible for the enhanced frequency of promoter deletion in pIAH011 plasmid in an evolved strain of PDLD phenotype **(Fig. 5.5)**. The recombination between the operators in the pIAH011 plasmid was established to occur at very low frequencies in the ancestral IAH015 and IAH016 strains (see Results I). However, in one of the evolved strains of PDLD phenotype, IAH016*, this mutation is occurring at an increased frequency **(Fig. 5.3).** It was previously discovered that *mutL* inactivity can promote RecA-independent recombination events, such as those observed in the pAVE011 and pIAH011 plasmids (Elez et al. 2007; Lovett & Feschenko 1996). Phenotypically, the promoter deletion in these plasmids corresponds to reduced productivity in presence and absence of the inducer. This promoter deletion mutation was the only plasmid mutation detected in the evolved groups of this experiment at a relevant frequency.

To further investigate the effect of the three DNA damage repair genes mutations, a statistical analysis was performed. The number of mutations (any type) per genome was recorded alongside the genotype of each clone. The analysed clones belonged to various phenotype groups (PULD, PDLD, EPEK, OPAK, EPAK, APAB*, APAK, NPEB*, NPEK). The summarised data is available in **Fig. 5.4.**



Total number of clones in each group	95	4	18	18	12	3	4

Fig. 5.4 The total number of mutations (any type) per genome in clones sequenced as part of the evolutionary experiment. The clones have been also classed by the presence of any mutation in one of the DNA damage repair genes *mutL, alkB* and *yoaA*. Displayed clones originated in various evolutionary phenotype and control groups (PULD, PDLD, EPEK, OPAK, EPAK, APAB*, APAK, NPEB*, NPEK).

The collected data was analysed using multiple linear regression model described below:

$$Y = \beta 0 + \beta 1 \times mutL + \beta 2 \times alkB + \beta 3 \times yoaA$$

Y is the number of mutations (any type) per genome.

β0 describes the model intercept, this is the average number of mutations (any type) per genome in *E. coli* clones involved in this study without mutations in either *mutL*, *alkB* or *yoaA* genes..

β1 describes the average increase in the number of mutations (any type) per genome in *E. coli* clones involved in this study with any mutation in *mutL*.

β2 describes the average increase in the number of mutations (any type) per genome in *E. coli* clones involved in this study with any mutation in *alkB*.

β3 describes the average increase in the number of mutations (any type) per genome in *E. coli* clones involved in this study with any mutation in *yoaA*.

According to the model, the average number of mutations per *E. coli* genome in this study (with no mutations in *mutL*, *alkB* or *yoaA* is 85 (95% CI 79.32 to 90.81). A mutation in either of these three genes increases this number. The degree of this increase is dependent on the gene affected by a mutation. For example, the clones with either *alkB* or *yoaA* mutation have an increased number of total mutations per genome by 10.16 and 13.54, respectively (95% CI 1.875 to 18.44; p=0.0168 and 5.327 to 21.76; p= 0.0015). However, any mutation affecting *mutL* increases the total number of mutations per genome by 29.33 (95% CI 16.93 to 41.73; p <0.0001). This confirms the hypothesis linking the increased number of mutations observed in some of the plasmids with the *mutL* mutations present on the host genome.

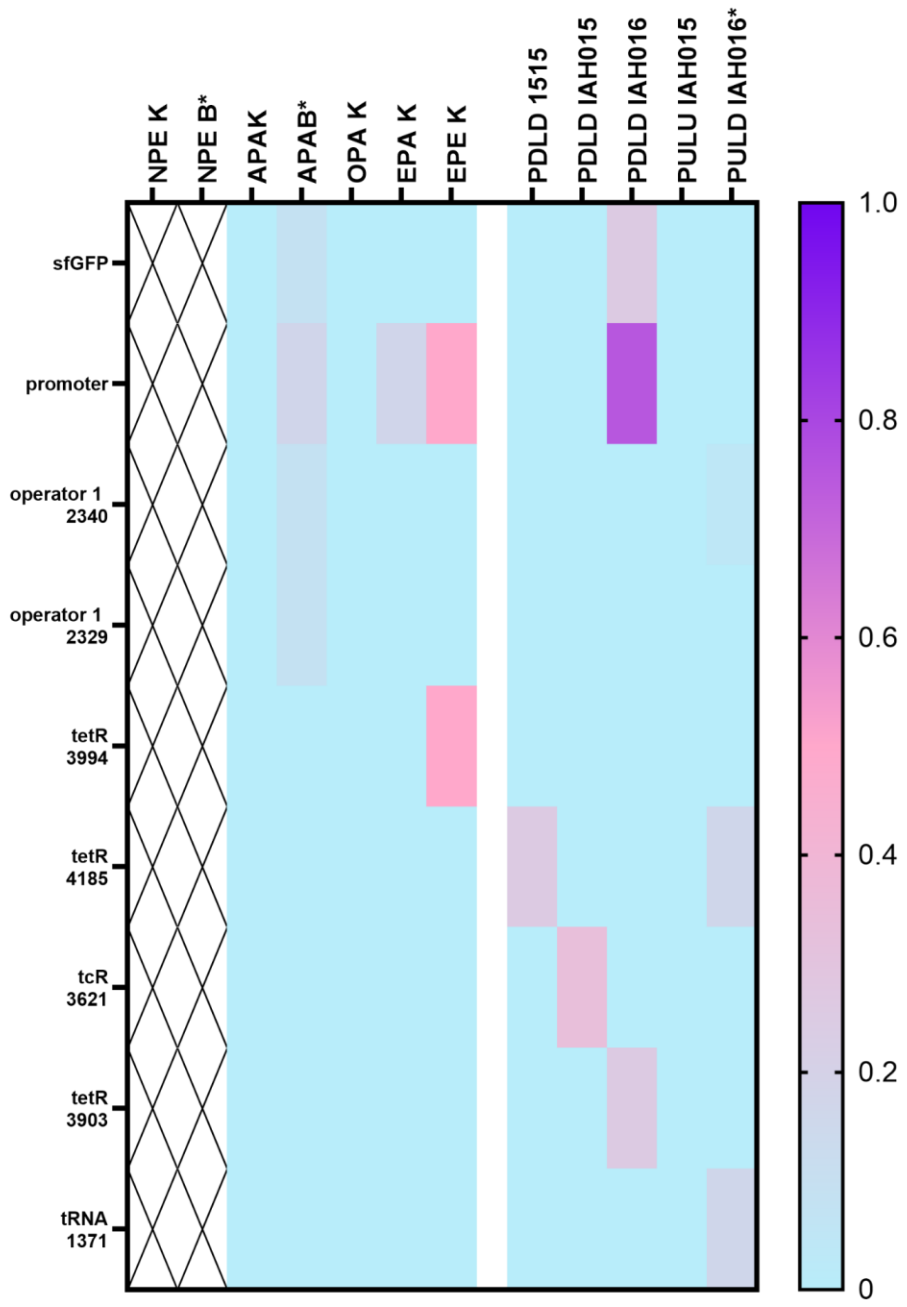


Fig. 5.5 The frequency of genetic changes affecting regions of the pAVEway plasmids in several phenotypic groups when compared to an ancestral plasmid map reference. Strains labelled with a * have been labelled as BL21 throughout the experiment, but have been now identified as W3110 genotype. The scale represents the fraction of each group where a polymorphism was identified in a given region; 0=present in no clones in a group, 1=present in all clones in a group. Group sizes were varied: altered plasmid ancestral K background (APAK) = 6 clones; altered plasmid ancestral B* background (APAB*) = 12 clones; original plasmid K background (OPAK)

= 6 clones; empty plasmid ancestral K background (EPAK) = 6 clones; empty plasmid evolved K background (EPEK) = 6 clones; productivity down leakiness down (PDLD) 1515 = 4 clones; PDLD IAH015 = 12 clones; PDLD IAH016* = 8 clones; productivity up leakiness up (PULU) IAH015 = 7 clones; productivity up leakiness down IAH016 = 25 clones.

5.2.3.4.2.2 DNA binding genes

The next three genes affected in the PDLD phenotype group of the evolved strains are encoding DNA-binding regulators. The first gene, *fhlA,* activates the transcription of the genes of the formate hydrogen lyase system. The mutation uncovered in the PDLD phenotype is a 3 bp deletion of the coding sequence resulting in Lysine 526 being removed from the protein product. Amino acids 379–693 of the FhlA protein are involved in its ATPase activity, as well as binding of the σ54 factor. Therefore, this mutation may impact the ability of FhlA to perform its function of activating transcription of the formate hydrogen lyase genes. Those genes are usually transcribed during anaerobic growth (Steinhilper et al. 2022) and since the experimental evolution conditions were aerobic, this is unlikely to be influencing the PDLD phenotype.

The second transcription regulator gene identified as a mutation hotspot in the PDLD phenotype is *rcsA.* The identified 4 bp deletion causes a frameshift and introduces a premature stop codon into the sequence. RcsA is a part of a regulatory network controlling the expression of capsular genes in response to osmotic shock (Sledjeski & Gottesman 1996). However, RcsA acts as a secondary activator in the system, together with RcsB for maximal capsule protein expression; in absence of RcsA, RcsB can still activate the transcription of the capsular genes (Gottesman & Stout 1991). This mutation's effect on the PDLD phenotype is therefore difficult to predict without validation experiment involving constructed mutants.

The final DNA-binding protein affected by the mutations in the PDLD phenotype group is Rob, however the uncovered mutation (G->A at the 778bp position of the coding sequence) is synonymous and has no effect on the protein functionality.

5.2.3.4.2.3 Transporter genes

There were two transporter genes affected in the PDLD phenotype group. The first one, *yjhB* (also known as *nanX*), carried a frameshift mutation caused by 1 bp deletion of an A nucleotide in a homopolymer A stretch of the DNA coding sequence. Considering the nanopore sequencing technology potential for mistakes in base calling in such regions, this mutation is unlikely to be of any significance to the PDLD phenotype. In the other affected gene, *mdtF,* the identified mutation was a 3 bp deletion resulting in Lysine 768 amino acid residue being removed from the protein product. This mutation falls between the 7th and 8th hypothesised transmembrane domains of the protein. Currently, the MdtF protein structure remains unresolved, and only an AlphaFold prediction is available (UniProt accession P37637). The secondary structure predicted by AlphaFold does not provide any information about potential sites crucial to MdtF activity as a multidrug efflux pump. Therefore, it is impossible to hypothesise about the impact of the uncovered mutation on the PDLD phenotype without further experiments.

5.2.3.4.3 Productivity up, leakiness down phenotype (PDLD)

This evolved phenotype was the most promising in terms of the application in the recombinant protein production industry. However, there were no gene coding regions which were uniquely affected in this group either in the genome **(Fig. 5.1)** or the in the plasmid **(Fig. 5.4)**

There are several explanations for this finding. Firstly, the PULD genotype shares commonalities with both PDLD and PULU groups **(Fig. 5.1).** It is possible that it is the unique combination of specific genes affected by mutations which is responsible for this phenotype. For example, *pcnB* mutation present in almost half of the isolated clones of PULD phenotype is the same mutation present in the PDLD phenotype group. It is therefore likely at least partially responsible for the decreased leakiness in both phenotypes. There are also genes affected by mutations in the PDLD group which are unaffected in the PULD group, highlighting the role of those mutations in maintaining the production of sfGFP low in the PDLD group.

Secondly, a number of non-coding regions have been identified as carrying mutations implicated in the various evolved phenotypes **(Fig. 5.5).** Those regions may be involved in regulating the neighbouring gene expression, for example if those non-coding regions contain operator or promoter sequences. The implications of those mutations are much more difficult to predict; nevertheless they should be investigated further as potential factors influencing the altered phenotypes of the evolved *E. coli* strains carrying pAVEway plasmids.

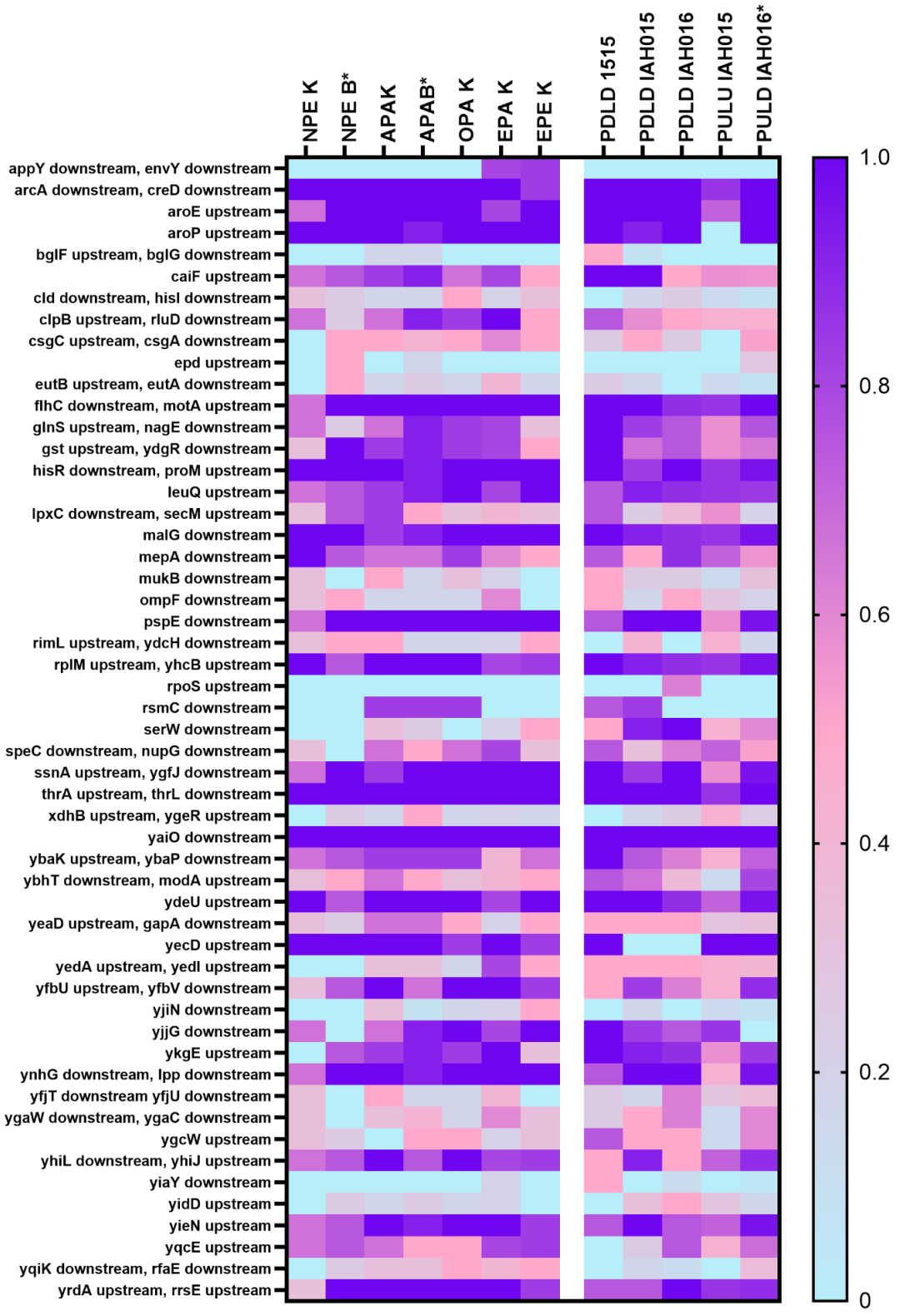


Fig. 5.6 The frequency of genetic changes affecting specific non-coding regions of the *E. coli* genome in several phenotypic groups when compared to no plasmid ancestor reference genome. Strains labelled with a * have been labelled as

BL21 throughout the experiment, but have been now identified as W3110 genotype. The scale represents the fraction of each group where a polymorphism was identified in a given region; 0=present in no clones in a group, 1=present in all clones in a group. Group sizes were varied: no plasmid evolved K background (NPEK) = 3 clones; no plasmid evolved B* background (NPEB*) = 4 clones; altered plasmid ancestral K background (APAK) = 6 clones; altered plasmid ancestral B* background (APAB*) = 12 clones; original plasmid K background (OPAK) = 6 clones; empty plasmid ancestral K background (EPAK) = 6 clones; empty plasmid evolved K background (EPEK) = 6 clones; productivity down leakiness down (PDLD) 1515 = 4 clones; PDLD IAH015 = 12 clones; PDLD IAH016* = 8 clones; productivity up leakiness up (PULU) IAH015 = 7 clones; productivity up leakiness down IAH016 = 25 clones.

6. Conclusions and future work

The work presented in this thesis has several implications for the recombinant protein production industry.

In “Results I” pAVEway plasmids were characterised. It was determined that pAVE011-carrying *E. coli* strains are sensitive to the stress imposed by heterologous protein expression induction and that induction in crucial growth phases (such as lag phase) results in increased rates of homologous recombination in the plasmid promoter region. These results are consistent with existing literature (James *et al.* 2021) and highlight the importance of careful manufacturing process design. While the exact mechanism of recombination remains undetermined, it was shown to be RecA-independent. RecA deficient strains have been previously identified as superior in terms of plasmid stability (Phue et al. 2008); however the results reported in “Results I” show that this benefit is dependent on plasmid design. pAVEway plasmids are recombination-prone due to the palindromic *lac* operators and are not stabilised in *recA*- host background.

This promoter recombination reduces, but not completely eliminates, the heterologous protein production. The observed productivity reduction is associated with the formation of a new promoter region within the recombined plasmids, which is weaker than the original promoter sequence. Replacing the operator sequence upstream of the plasmid promoter with an alternative palindromic sequence was sufficient to significantly reduce the frequency of homologous recombination. Further investigations have revealed that homologous recombination is not the only mechanism responsible for reduced culture productivity in strains which were previously exposed to inducer stress. This observation of reduced productivity in multiple induction cycles has been previously observed in recombinant protein

producing strains and reported as hindrance in continuous fermentation process development (Sieben *et al.* 2016; Vyas *et al.* 1994).

The pAVEway strain responses to various media components have also been investigated, as all strains transformed with either pAVE011 or pIAH011 plasmid have shown “leaky” protein expression in the absence of IPTG induction. This was shown to be linked to starvation response and is mediated by glucose supplementation. Glucose supplementation may therefore be used at scale to prevent “leaky expression” of products potentially toxic to *E. coli.*

In “Results II” a novel FACS-assisted screen was described which can be used to direct experimental evolution of heterologous protein producing *E. coli* towards a more stable and more productive phenotype. One of the main challenges of this method was preventing cross-contamination between the different culture strains and lineages. The described steps undertaken to prevent this issue (FACS machine cleaning, experimental design separating the BL21 and W3110 background strains as well as employment of an original plate lid design during cell sorting steps) were ultimately insufficient to completely prevent mixing of the cultures. However, given the uniquely specific direction of contamination (always from W3110 background cultures to BL21, from pIAH011 cultures to pAVE011 cultures) the source of the contamination was identified as inadequate FACS machine cleaning between cultures of different strains. In the future use in experimental evolution protocols, it is recommended that only samples of the same culture genotype are sorted on the same day, and a rigorous cleaning protocol is employed before attempting to sort samples of a different genotype. In other experimental designs, where avoiding cross-contamination between replicates of the same genotype is required, the samples should be sorted

on different days, allowing for a full cleaning protocol of the equipment to be completed between samples.

Despite its challenges, the FACS-based method yielded the desired results. Clones evolved over 7 weeks and phenotypically screened against their ancestors exhibited a wide variety of phenotypes, some of which were potentially relevant to the heterologous protein production industry. One specific phenotype with higher productivity and lower "leaky expression" was identified (PULD), which exhibited up to 3.3 times higher GFP expression under induced conditions, up to 17.7 times lower GFP expression in absence of inducer, and improved stability.

Finally, in "Results III" some regions of interest were identified through a high-throughput sequencing data analysis as mutation targets responsible for the altered evolved phenotypes. One such target, the *pcnB* coding region can be responsible for one of the two opposing phenotypes - productivity up, leakiness up (PULU) and productivity down, leakiness down (PULD). The phenotype presentation is dependent on the frameshift mutation location in either the N-terminal domain, crucial to PcnB function, or the C- terminal domain, which may be dispensable. Future work could experimentally investigate the effects of these mutations on the PcnB function.

Multiple intergenic regions were also identified as potential mutational hotspots playing a role in phenotype presentation of the evolved *E. coli* strains. Those also can be further investigated via combination of bioinformatics and experimental work.

Several questions remain unanswered by the results presented here. More experimental investigations are needed to understand if the uncovered genetic variants associated with evolved phenotypes, especially PULD, are observed in strains carrying plasmids encoding other target proteins. Furthermore, the variants should be cloned alone and in combination into unevolved *E. coli* to validate their

function as underlying causes of the phenotypes. Finally, experiments investigating the PULD clones in bioreactor environment should be conducted to verify that the observed phenotype improvement is stable, and production using those clones can be scaled up.

7. Reference list

Abushaheen, M.A. et al., 2020. Antimicrobial resistance, mechanisms and its clinical significance. *Disease-a-month: DM*, 66(6). Available at: https://pubmed.ncbi.nlm.nih.gov/32201008/ [Accessed February 13, 2024].

Aidelberg, G. et al., 2014. Hierarchy of non-glucose sugars in Escherichia coli. *BMC systems biology*, 8(1), pp.1–12.

Al-Hawash, A.B., Zhang, X. & Ma, F., 2017. Strategies of codon optimization for high-level heterologous protein expression in microbial expression systems. *Gene Reports*, 9, pp.46–53.

Allen, J.R. et al., 2022. Segregationally stabilised plasmids improve production of commodity chemicals in glucose-limited continuous fermentation. *Microbial cell factories*, 21(1), pp.1–11.

Alvarez-Sieiro, P. et al., 2014. Generation of food-grade recombinant Lactobacillus casei delivering Myxococcus xanthus prolyl endopeptidase. *Applied microbiology and biotechnology*, 98(15), pp.6689–6700.

Anindyajati et al., 2016. Plasmid Copy Number Determination by Quantitative Polymerase Chain Reaction. *Scientia pharmaceutica*, 84(1), p.89.

Arigoni, F. et al., 1998. A genome-based approach for the identification of essential bacterial genes. *Nature biotechnology*, 16(9). Available at: https://pubmed.ncbi.nlm.nih.gov/9743119/ [Accessed April 28, 2024].

Baba, T. et al., 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology*, 2, p.2006.0008.

Badía, J. et al., 1998. A rare 920-kilobase chromosomal inversion mediated by IS1 transposition causes constitutive expression of the yiaK-S operon for carbohydrate utilization in Escherichia coli. *The Journal of biological chemistry*, 273(14). Available at: https://pubmed.ncbi.nlm.nih.gov/9525947/ [Accessed April 28, 2024].

Balleza, E., Mark Kim, J. & Cluzel, P., 2018. Systematic characterization of maturation time of fluorescent proteins in living cells. *Nature methods*, 15(1), p.47.

Banasik, M. & Sachadyn, P., 2014. Conserved motifs of MutL proteins. *Mutation research*, 769. Available at: https://pubmed.ncbi.nlm.nih.gov/25771726/ [Accessed April 29, 2024].

Ban, C. & Yang, W., 1998. Crystal Structure and ATPase Activity of MutL: Implications for DNA Repair and Mutagenesis. *Cell*, 95(4), pp.541–552.

Baneyx, F. & Georgiou, G., 1990. In vivo degradation of secreted fusion proteins by the Escherichia coli outer membrane protease OmpT. *Journal of bacteriology*. Available at: https://journals.asm.org/doi/10.1128/jb.172.1.491-494.1990 [Accessed February 12, 2024].

Bartolo-Aguilar, Y. et al., 2022. The potential of cold-shock promoters for the expression of recombinant proteins in microbes and mammalian cells. *Journal of Genetic Engineering and Biotechnology*, 20(1), pp.1–20.

Bashir, A. et al., 1995. Altering kinetic mechanism and enzyme stability by mutagenesis of the dimer interface of glutathione reductase. *Biochemical Journal*, 312(2), pp.527–533.

Basit, A. et al., 2018. Health improvement of human hair and their reshaping using recombinant keratin K31. *Biotechnology Reports*, 20, p.e00288.

Bazaral, M. & Helinski, D.R., 1968. Circular DNA forms of colicinogenic factors E1, E2 and E3 from Escherichia coli. *Journal of molecular biology*, 36(2). Available at: https://pubmed.ncbi.nlm.nih.gov/4939624/ [Accessed February 12, 2024].

Behrens, F. et al., 2015. MOR103, a human monoclonal antibody to granulocyte–macrophage colony-stimulating factor, in the treatment of patients with moderate rheumatoid arthritis: results of a phase Ib/IIa randomised, double-blind, placebo-controlled, dose-escalation trial. *Annals of the rheumatic diseases*, 74(6), pp.1058–1064.

Belasco, J.G. et al., 1986. The stability of E. coli gene transcripts is dependent on determinants localized to specific mRNA segments. *Cell*, 46(2), pp.245–251.

Benz, F. & Hall, A.R., 2023. Host-specific plasmid evolution explains the variable spread of clinical antibiotic-resistance plasmids. *Proceedings of the National Academy of Sciences of the United States of America*, 120(15). Available at: https://pubmed.ncbi.nlm.nih.gov/37023131/ [Accessed February 13, 2024].

Berthold, H. et al., 1992. Plasmid pGEX-5T: An alternative system for expression and purification of recombinant proteins. *Biotechnology letters*, 14(4), pp.245–250.

Bertrand, R.L., 2019. Lag Phase Is a Dynamic, Organized, Adaptive, and Evolvable Period That Prepares Bacteria for Cell Division. *Journal of bacteriology*, 201(7). Available at: http://dx.doi.org/10.1128/JB.00697-18.

Bhagwat, A.S. & Person, S., 1981. Structure and properties of the region of homology between plasmids pMB1 and ColE1. *Molecular & general genetics: MGG*, 182(3), pp.505–507.

Bi, X. & Liu, L.F., 1994. recA-independent and recA-dependent Intramolecular Plasmid Recombination: Differential Homology Requirement and Distance Effect. *Journal of molecular biology*, 235(2), pp.414–423.

Boël, G. et al., 2016. Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature*, 529(7586), pp.358–363.

de Boer, H.A., Comstock, L.J. & Vasser, M., 1983. The tac promoter: a functional hybrid derived from the trp and lac promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 80(1), p.21.

Bolivar, F., Rodriguez, R.L., Betlach, M.C., et al., 1977. Construction and characterization of new cloning vehicles. I. Ampicillin-resistant derivatives of the plasmid pMB9. *Gene*, 2(2). Available at: https://pubmed.ncbi.nlm.nih.gov/344136/ [Accessed February 12, 2024].

Bolivar, F., Rodriguez, R.L., Greene, P.J., et al., 1977. Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. *Gene*, 2(2). Available at: https://pubmed.ncbi.nlm.nih.gov/344137/ [Accessed February 12, 2024].

Bottery, M.J., Wood, A.J. & Brockhurst, M.A., 2017. Adaptive modulation of antibiotic resistance through intragenomic coevolution. *Nature Ecology & Evolution*, 1(9), pp.1364–1369.

Bottery, M.J., Wood, A.J. & Brockhurst, M.A., 2018. Temporal dynamics of bacteria-plasmid coevolution under antibiotic selection. *The ISME journal*, 13(2), pp.559–562.

Brämer, C. et al., 2019. Optimization of continuous purification of recombinant patchoulol synthase from Escherichia coli with membrane adsorbers. *Biotechnology progress*, 35(4), p.e2812.

Brooks, S.A., 2004. Appropriate glycosylation of recombinant proteins for human use. *Molecular biotechnology*, 28(3), pp.241–255.

Brown, L.T. et al., 2015. Connecting Replication and Repair: YoaA, a Helicase-Related Protein, Promotes Azidothymidine Tolerance through Association with Chi, an Accessory Clamp Loader Protein. *PLoS genetics*, 11(11). Available at: https://pubmed.ncbi.nlm.nih.gov/26544712/ [Accessed April 28, 2024].

Bryant, F.R., 1988. Construction of a recombinase-deficient mutant recA protein that retains single-stranded DNA-dependent ATPase activity. *The Journal of biological chemistry*, 263(18). Available at: https://pubmed.ncbi.nlm.nih.gov/2967815/ [Accessed November 18, 2021].

Buchet, A. et al., 1999. Positive co-regulation of the Escherichia coli carnitine pathway cai and fix operons by CRP and the CaiF activator. *Molecular microbiology*, 34(3). Available at: https://pubmed.ncbi.nlm.nih.gov/10564497/ [Accessed April 28, 2024].

Buckholz, R.G. & Gleeson, M.A.G., 1991. Yeast Systems for the Commercial Production of Heterologous Proteins. *Biotechnology* , 9(11), pp.1067–1072.

Burgess-Brown, N.A. et al., 2008. Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study. *Protein expression and purification*, 59(1). Available at: https://pubmed.ncbi.nlm.nih.gov/18289875/ [Accessed February 12, 2024].

Bzymek, M. & Lovett, S.T., 2001. Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. *Proceedings of the National Academy of Sciences of the United States of America*, 98(15), pp.8319–8325.

Campos, N. et al., 2001. Escherichia coli engineered to synthesize isopentenyl diphosphate and dimethylallyl diphosphate from mevalonate: a novel system for the genetic analysis of the 2-C-methyl-d-erythritol 4-phosphate pathway for isoprenoid biosynthesis. *The Biochemical journal*, 353(Pt 1). Available at: https://pubmed.ncbi.nlm.nih.gov/11115399/ [Accessed April 27, 2024].

Carrier, T., Jones, K.L. & Keasling, J.D., 1998. mRNA stability and plasmid copy number effects on gene expression from an inducible promoter system. *Biotechnology and bioengineering*, 59(6), pp.666–672.

Carroll, A.C. & Wong, A., 2018. Plasmid persistence: costs, benefits, and the plasmid paradox. *Canadian journal of microbiology*. Available at: https://cdnsciencepub.com/doi/full/10.1139/cjm-2017-0609 [Accessed April 11, 2022].

Çelik, E. & Çalık, P., 2012. Production of recombinant proteins by yeast cells. *Biotechnology advances*, 30(5), pp.1108–1118.

Cesareni, G., Helmer-Citterich, M. & Castagnoli, L., 1991. Control of ColE1 plasmid replication by antisense RNA. *Trends in genetics: TIG*, 7(7), pp.230–235.

Chang, A.C.Y. et al., 1978. Phenotypic expression in E. coli of a DNA sequence coding for mouse dihydrofolate reductase. *Nature*, 275(5681), pp.617–624.

Chédin, F. et al., 1994. Frequency of deletion formation decreases exponentially with distance between short direct repeats. *Molecular microbiology*, 12(4). Available at: https://pubmed.ncbi.nlm.nih.gov/7934879/ [Accessed April 14, 2022].

Chen, G.Q. & Jiang, X.R., 2018. Next generation industrial biotechnology based on extremophilic bacteria. *Current opinion in biotechnology*, 50. Available at: https://pubmed.ncbi.nlm.nih.gov/29223022/ [Accessed March 21, 2024].

Choi, J.H. et al., 2000. Efficient secretory production of alkaline phosphatase by high cell density culture of recombinant Escherichia coli using the Bacillus sp. endoxylanase signal sequence. *Applied microbiology and biotechnology*, 53(6), pp.640–645.

Chou, C.H. et al., 1995. Characterization of a pH-inducible promoter system for high-level expression of recombinant proteins in Escherichia coli. *Biotechnology and bioengineering*, 47(2), pp.186–192.

Coburn, G.A. et al., 1999. Reconstitution of a minimal RNA degradosome demonstrates functional coordination between a 3' exonuclease and a DEAD-box RNA helicase. *Genes & development*, 13(19). Available at: https://pubmed.ncbi.nlm.nih.gov/10521403/ [Accessed April 28, 2024].

Cohen, S.N. et al., 1973. Construction of Biologically Functional Bacterial Plasmids In Vitro. *Proceedings of the National Academy of Sciences (PNAS)*, 70(11), pp.3240–32444.

Cohen, S.N. & Chang, A.C.Y., 1973. Recircularization and Autonomous Replication of a Sheared R-Factor DNA Segment in Escherichia coli Transformants. *Proceedings of the National Academy of Sciences of the United States of America*, 70(5), p.1293.

Cranenburgh, R.M., Lewis, K.S. & Hanak, J.A.J., 2004. Effect of Plasmid Copy Number and lac Operator Sequence on Antibiotic-Free Plasmid Selection by Operator-Repressor Titration in Escherichia coli. *Journal of molecular microbiology and biotechnology*, 7(4), pp.197–203.

Csörgő, B. et al., 2012. Low-mutation-rate, reduced-genome Escherichia coli: an improved host for faithful maintenance of engineered genetic constructs. *Microbial cell factories*, 11(1), pp.1–13.

Daegelen, P. et al., 2009. Tracing ancestors and relatives of Escherichia coli B, and the derivation of B strains REL606 and BL21(DE3). *Journal of molecular biology*, 394(4). Available at: https://pubmed.ncbi.nlm.nih.gov/19765591/ [Accessed February 12, 2024].

Dassain, M. et al., 1999. A new essential gene of the "minimal genome" affecting cell division. *Biochimie*, 81(8-9). Available at: https://pubmed.ncbi.nlm.nih.gov/10572302/ [Accessed April 28, 2024].

Datsenko, K.A. & Wanner, B.L., 2000. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences of the United States of America*, 97(12), pp.6640–6645.

Deatherage, D.E. et al., 2018. Directed evolution of Escherichia coli with lower-than-natural plasmid mutation rates. *Nucleic acids research*, 46(17), pp.9236–9250.

De Gelder, L. et al., 2008. Adaptive Plasmid Evolution Results in Host-Range Expansion of a Broad-Host-Range Plasmid. *Genetics*, 178(4), pp.2179–2190.

Delahaye, C. & Nicolas, J., 2021. Sequencing DNA with nanopores: Troubles and biases. *PloS one*, 16(10), p.e0257521.

Demain, A.L. & Vaishnav, P., 2009. Production of recombinant proteins by microbes and higher organisms. *Biotechnology advances*, 27(3). Available at: https://pubmed.ncbi.nlm.nih.gov/19500547/ [Accessed January 16, 2024].

Ding, N. et al., 2017. Increased glycosylation efficiency of recombinant proteins in Escherichia coli by auto-induction. *Biochemical and biophysical research communications*, 485(1). Available at: https://pubmed.ncbi.nlm.nih.gov/28188786/ [Accessed February 7, 2024].

Du, F. et al., 2021. Regulating the T7 RNA polymerase expression in E. coli BL21 (DE3) to provide more host options for recombinant protein production. *Microbial cell factories*, 20(1), pp.1–10.

Dürrschmid, K. et al., 2008. Monitoring of transcriptome and proteome profiles to investigate the cellular response of E. coli towards recombinant protein expression under defined chemostat conditions. *Journal of biotechnology*, 135(1), pp.34–44.

Eichler, K. et al., 1996. Identification and characterization of the caiF gene encoding a potential transcriptional activator of carnitine metabolism in Escherichia coli. *Journal of bacteriology*, 178(5). Available at: https://pubmed.ncbi.nlm.nih.gov/8631699/ [Accessed April 28, 2024].

Elez, M., Radman, M. & Matic, I., 2007. The frequency and structure of recombinant products is determined by the cellular level of MutL. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21). Available at: https://pubmed.ncbi.nlm.nih.gov/17502621/ [Accessed April 29, 2024].

Esposito, D. & Chatterjee, D.K., 2006. Enhancement of soluble protein expression through the use of fusion tags. *Current opinion in biotechnology*, 17(4). Available at: https://pubmed.ncbi.nlm.nih.gov/16781139/ [Accessed February 12, 2024].

Eurofins Genomics, Sequencing Primer Design Tool. Available at: https://eurofinsgenomics.eu/en/ecom/tools/sequencing-primer-design/ [Accessed February 5, 2024].

Feldman, M.F. et al., 2005. Engineering N-linked protein glycosylation with diverse O antigen lipopolysaccharide structures in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8). Available at: https://pubmed.ncbi.nlm.nih.gov/15703289/ [Accessed February 7, 2024].

Field, C.M. & Summers, D.K., 2011. Multicopy plasmid stability: revisiting the dimer catastrophe. *Journal of theoretical biology*, 291. Available at: https://pubmed.ncbi.nlm.nih.gov/21945338/ [Accessed May 16, 2022].

Finn, R.D. et al., 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic acids research*, 44(D1). Available at: https://pubmed.ncbi.nlm.nih.gov/26673716/ [Accessed April 27, 2024].

Fordjour, E. et al., 2023. Improved Membrane Permeability via Hypervesiculation for In Situ Recovery of Lycopene in Escherichia coli. *ACS synthetic biology*. Available at: https://pubs.acs.org/doi/full/10.1021/acssynbio.3c00306 [Accessed February 7, 2024].

Foster, P.L. et al., 2015. Determinants of spontaneous mutation in the bacterium Escherichia coli as revealed by whole-genome sequencing. *Proceedings of the National Academy of Sciences*, 112(44), pp.E5990–E5999.

Freiberg, C. et al., 2001. Identification of novel essential Escherichia coli genes conserved among pathogenic bacteria. *Journal of molecular microbiology and biotechnology*, 3(3). Available at: https://pubmed.ncbi.nlm.nih.gov/11361082/ [Accessed April 27, 2024].

Fric, J. et al., 2012. Use of Human Monoclonal Antibodies to Treat Chikungunya Virus Infection. *The Journal of infectious diseases*, 207(2), pp.319–322.

Furrer, J.L. et al., 2002. Export of the siderophore enterobactin in Escherichia coli: involvement of a 43 kDa membrane exporter. *Molecular microbiology*, 44(5). Available at: https://pubmed.ncbi.nlm.nih.gov/12068807/ [Accessed April 27, 2024].

Gaciarz, A. et al., 2016. Systematic screening of soluble expression of antibody fragments in the cytoplasm of E. coli. *Microbial cell factories*, 15(1), pp.1–10.

Gerdes, S.Y. et al., 2003. Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *Journal of bacteriology*, 185(19). Available at: https://pubmed.ncbi.nlm.nih.gov/13129938/ [Accessed April 28, 2024].

Glasscock, C.J. et al., 2018. A flow cytometric approach to engineering Escherichia coli for improved eukaryotic protein glycosylation. *Metabolic engineering*, 47. Available at: https://pubmed.ncbi.nlm.nih.gov/29702274/ [Accessed February 7, 2024].

Goeddel, D.V. et al., 1979a. Expression in Escherichia coli of chemically synthesized genes for human insulin. *Proceedings of the National Academy of Sciences of the United States of America*, 76(1), pp.106–110.

Goeddel, D.V. et al., 1979b. Expression in Escherichia coli of chemically synthesized genes for human insulin. *Proceedings of the National Academy of Sciences (PNAS)*, 76(1), pp.106–110.

Goodall, E.C.A. et al., 2018. The Essential Genome of Escherichia coli K-12. *mBio*, 9(1). Available at: https://pubmed.ncbi.nlm.nih.gov/29463657/ [Accessed April 28, 2024].

Gottesman, S. & Stout, V., 1991. Regulation of capsular polysaccharide synthesis in Escherichia coli K12. *Molecular microbiology*, 5(7). Available at: https://pubmed.ncbi.nlm.nih.gov/1943696/ [Accessed April 29, 2024].

Govender, K. et al., 2020. A novel and more efficient biosynthesis approach for human insulin production in Escherichia coli (E. coli). *AMB Express*, 10(1), pp.1–9.

Grant, S.G. et al., 1990. Differential plasmid rescue from transgenic mouse DNAs into Escherichia coli methylation-restriction mutants. *Proceedings of the National Academy of Sciences of the United States of America*, 87(12). Available at: https://pubmed.ncbi.nlm.nih.gov/2162051/ [Accessed February 13, 2024].

Grossman, T.H. et al., 1998. Spontaneous cAMP-dependent derepression of gene expression in stationary phase plays a role in recombinant expression instability. *Gene*, 209(1-2), pp.95–103.

Guan, K.L. & Dixon, J.E., 1991. Eukaryotic proteins expressed in Escherichia coli: an improved thrombin cleavage and purification procedure of fusion proteins with glutathione S-transferase. *Analytical biochemistry*, 192(2). Available at: https://pubmed.ncbi.nlm.nih.gov/1852137/ [Accessed February 12, 2024].

Gulmez, C. et al., 2018. A novel detergent additive: Organic solvent- and thermo-alkaline-stable recombinant subtilisin. *International journal of biological macromolecules*, 108, pp.436–443.

Gundinger, T. et al., 2022. Recombinant Protein Production in E. coli Using the phoA Expression System. *Fermentation*, 8(4), p.181.

Hakes, D.J. & Dixon, J.E., 1992. New vectors for high level expression of recombinant proteins in bacteria. *Analytical biochemistry*, 202(2). Available at: https://pubmed.ncbi.nlm.nih.gov/1519755/ [Accessed February 12, 2024].

Hall, J.P.J. et al., 2021. Plasmid fitness costs are caused by specific genetic conflicts enabling resolution by compensatory mutation. *PLoS biology*, 19(10), p.e3001225.

Hanahan, D., 1983. Studies on transformation of Escherichia coli with plasmids. *Journal of molecular biology*, 166(4). Available at: https://pubmed.ncbi.nlm.nih.gov/6345791/ [Accessed February 13, 2024].

Han, M.-J. et al., 2004. Roles and applications of small heat shock proteins in the production of recombinant proteins in Escherichia coli. *Biotechnology and bioengineering*, 88(4), pp.426–436.

Hausjell, J. et al., 2018. E. coli HMS174(DE3) is a sustainable alternative to BL21(DE3). *Microbial cell factories*, 17(1). Available at: https://pubmed.ncbi.nlm.nih.gov/30376846/ [Accessed February 13, 2024].

Heim, R. & Tsien, R.Y., 1996. Engineering green fluorescent protein for improved brightness, longer wavelengths and fluorescence resonance energy transfer. *Current biology: CB*, 6(2), pp.178–182.

He, L. et al., 1993. PcnB is required for the rapid degradation of RNAI, the antisense RNA that controls the copy number of CoIE1-related plasmids. *Molecular microbiology*, 9(6), pp.1131–1142.

Herrera, G. et al., 2002. Assessment of Escherichia coli B with enhanced permeability to fluorochromes for flow cytometric assays of bacterial cell function. *Cytometry*, 49(2). Available at: https://pubmed.ncbi.nlm.nih.gov/12357461/ [Accessed January 17, 2024].

Hershfield, V. et al., 1974. Plasmid ColE1 as a Molecular Vehicle for Cloning and Amplification of DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 71(9), p.3455.

Hewitt, C.J. et al., 1999. The use of multi-parameter flow cytometry to compare the physiological response of Escherichia coli W3110 to glucose limitation during batch, fed-batch and continuous culture cultivations. *Journal of biotechnology*, 75(2-3), pp.251–264.

Hitzeman, R.A. et al., 1981. Expression of a human gene for interferon in yeast. *Nature*, 293(5835), pp.717–722.

Hodgson, I.J., Lennon, C.D.J. & Kara, B.V., 2013. EP2386642B1 Expression system. *European Patent*. Available at: http://dx.doi.org/10.1073/PNAS.85.23.8973.

Huemer, M. et al., 2020. Antibiotic resistance and persistence-Implications for human health and treatment perspectives. *EMBO reports*, 21(12). Available at: https://pubmed.ncbi.nlm.nih.gov/33400359/ [Accessed February 13, 2024].

Hughes, J.M. et al., 2012. The Role of Clonal Interference in the Evolutionary Dynamics of Plasmid-Host Adaptation. *mBio*. Available at: https://journals.asm.org/doi/10.1128/mbio.00077-12 [Accessed March 7, 2024].

Hülter, N.F. et al., 2020. Intracellular Competitions Reveal Determinants of Plasmid Evolutionary Success. *Frontiers in microbiology*, 11, p.558594.

Hu, T. et al., 2021. Next-generation sequencing technologies: An overview. *Human immunology*, 82(11). Available at: https://pubmed.ncbi.nlm.nih.gov/33745759/ [Accessed April 26, 2024].

Itoh, T. & Tomizawa, J., 1980. Formation of an RNA primer for initiation of replication of ColE1 DNA by ribonuclease H. *Proceedings of the National Academy of Sciences of the United States of America*, 77(5), pp.2450–2454.

Jagessar, K.L. & Jain, C., 2010. Functional and molecular analysis of Escherichia coli strains lacking multiple DEAD-box helicases. *RNA* , 16(7). Available at: https://pubmed.ncbi.nlm.nih.gov/20484467/ [Accessed April 28, 2024].

James, J. et al., 2021. Protein over-expression in Escherichia coli triggers adaptation analogous to antimicrobial resistance. *Microbial cell factories*, 20(1), p.13.

Johan H. J. Leveau, S.E.L., 2001. Predictive and Interpretive Simulation of Green Fluorescent Protein Expression in Reporter Bacteria. *Journal of bacteriology*, 183(23), p.6752.

Johnson, D.B. et al., 2012. Release factor one is nonessential in Escherichia coli. *ACS chemical biology*, 7(8). Available at: https://pubmed.ncbi.nlm.nih.gov/22662873/ [Accessed April 28, 2024].

Jones, K.L., Kim, S.W. & Keasling, J.D., 2000. Low-copy plasmids can perform as well as or better than high-copy plasmids for metabolic engineering of bacteria. *Metabolic engineering*, 2(4), pp.328–338.

Jordt, H. et al., 2020. Coevolution of host–plasmid pairs facilitates the emergence of novel multidrug resistance. *Nature Ecology & Evolution*, 4(6), pp.863–869.

Joshi, S.H.-N., Yong, C. & Gyorgy, A., 2022. Inducible plasmid copy number control for synthetic biology in commonly used E. coli strains. *Nature communications*, 13(1), pp.1–16.

Jurėnas, D. et al., 2022. Biology and evolution of bacterial toxin–antitoxin systems. *Nature reviews. Microbiology*, 20(6), pp.335–350.

Kanjee, U. & Houry, W.A., 2013. Mechanisms of acid resistance in Escherichia coli. *Annual review of microbiology*, 67. Available at: https://pubmed.ncbi.nlm.nih.gov/23701194/ [Accessed April 27, 2024].

Kawashima, H. et al., 1984. Functional domains of Escherichia coli recA protein deduced from the mutational sites in the gene. *Molecular and General Genetics MGG*, 193(2), pp.288–292. Available at: http://dx.doi.org/10.1007/bf00330682.

Khemici, V. et al., 2004. The RNase E of Escherichia coli has at least two binding sites for DEAD-box RNA helicases: functional replacement of RhlB by RhlE. *Molecular microbiology*, 54(5), pp.1422–1430.

Khmelinskii, A. et al., 2015. Incomplete proteasomal degradation of green fluorescent proteins in the context of tandem fluorescent protein timers. *Molecular biology of the cell*. Available at: https://www.molbiolcell.org/doi/full/10.1091/mbc.e15-07-0525 [Accessed April 28, 2022].

Kimura, S. & Suzuki, T., 2010. Fine-tuning of the ribosomal decoding center by conserved methyl-modifications in the Escherichia coli 16S rRNA. *Nucleic acids research*, 38(4). Available at: https://pubmed.ncbi.nlm.nih.gov/19965768/ [Accessed April 28, 2024].

Koita, K. & Rao, C.V., 2012. Identification and analysis of the putative pentose sugar efflux transporters in Escherichia coli. *PloS one*, 7(8). Available at: https://pubmed.ncbi.nlm.nih.gov/22952739/ [Accessed April 27, 2024].

Kopp, J. et al., 2020. Repetitive Fed-Batch: A Promising Process Mode for Biomanufacturing With E. coli. *Frontiers in Bioengineering and Biotechnology*, 8, p.573607.

Kosinski, J. et al., 2005. Analysis of the quaternary structure of the MutL C-terminal domain. *Journal of molecular biology*, 351(4). Available at: https://pubmed.ncbi.nlm.nih.gov/16024043/ [Accessed April 29, 2024].

Kroll, J. et al., 2009. Establishment of a novel anabolism-based addiction system with an artificially introduced mevalonate pathway: Complete stabilization of plasmids as universal application in white biotechnology. *Metabolic engineering*, 11(3), pp.168–177.

Kroll, J., Klinter, S., Schneider, C., et al., 2010. Plasmid addiction systems: perspectives and applications in biotechnology. *Microbial biotechnology*, 3(6), pp.634–657.

Kroll, J., Klinter, S. & Steinbüchel, A., 2010. A novel plasmid addiction system for large-scale production of cyanophycin in Escherichia coli using mineral salts medium. *Applied microbiology and biotechnology*, 89(3), pp.593–604.

Kuo, L.C. et al., 1988. Site-directed mutagenesis of Escherichia coli ornithine transcarbamoylase: role of arginine-57 in substrate binding and catalysis. *Biochemistry*, 27(24). Available at: https://pubmed.ncbi.nlm.nih.gov/3072022/ [Accessed April 27, 2024].

Lamrabet, O. et al., 2019. Changes in Intrinsic Antibiotic Susceptibility during a Long-Term Evolution Experiment with Escherichia coli. *mBio*. Available at: https://journals.asm.org/doi/10.1128/mbio.00189-19 [Accessed March 19, 2024].

Lee, H. et al., 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 109(41), pp.E2774–83.

Legrain, C., Stalon, V. & Glansdorff, N., 1976. Escherichia coli ornithine carbamolytransferase isoenzymes: evolutionary significance and the isolation of lambdaargF and lambdaargI transducing bacteriophages. *Journal of bacteriology*, 128(1). Available at: https://pubmed.ncbi.nlm.nih.gov/789338/ [Accessed April 27, 2024].

Lennon, C., 2014. White Paper: pAVEway™ expression system for the efficient expression of therapeutic proteins.

Lénon, M. et al., 2020. Improved production of Humira antibody in the genetically engineered Escherichia coli SHuffle, by co-expression of human PDI-GPx7 fusions. *Applied microbiology and biotechnology*, 104(22), pp.9693–9706.

Lenski, R.E. et al., 1991. Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. *The American naturalist*. Available at: https://www.journals.uchicago.edu/doi/10.1086/285289 [Accessed March 19, 2024].

Li, F. et al., 2010. Cell culture processes for monoclonal antibody production. *mAbs*. Available at: https://www.tandfonline.com/doi/abs/10.4161/mabs.2.5.12720 [Accessed February 7, 2024].

Lin-Chao, S. & Bremer, H., 1986. Effect of the bacterial growth rate on replication control of plasmid pBR322 in Escherichia coli. *Molecular & general genetics: MGG*, 203(1), pp.143–149.

Lin-Chao, S., Chen, W.T. & Wong, T.T., 1992. High copy number of the pUC plasmid results from a Rom/Rop-suppressible point mutation in RNA II. *Molecular microbiology*, 6(22). Available at: https://pubmed.ncbi.nlm.nih.gov/1283002/ [Accessed January 22, 2024].

Li, T. et al., 2014. Open and continuous fermentation: Products, conditions and bioprocess economy. *Biotechnology journal*, 9(12), pp.1503–1511.

Liu, J.D. & Parkinson, J.S., 1989. Genetics and sequence analysis of the pcnB locus, an Escherichia coli gene involved in plasmid copy number control. *Journal of bacteriology*. Available at: https://journals.asm.org/doi/10.1128/jb.171.3.1254-1261.1989 [Accessed April 28, 2024].

Liu, L. et al., 2020. Repeated fed-batch strategy and metabolomic analysis to achieve high docosahexaenoic acid productivity in Crypthecodinium cohnii. *Microbial cell factories*, 19(1), pp.1–14.

Liu, M.Y. et al., 1997. The RNA molecule CsrB binds to the global regulatory protein CsrA and antagonizes its activity in Escherichia coli. *The Journal of biological chemistry*, 272(28). Available at: https://pubmed.ncbi.nlm.nih.gov/9211896/ [Accessed April 27, 2024].

Lobstein, J. et al., 2012. SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. *Microbial cell factories*, 11(1), pp.1–16.

Loftie-Eaton, W. et al., 2015. Evolutionary Paths That Expand Plasmid Host-Range: Implications for Spread of Antibiotic Resistance. *Molecular biology and evolution*, 33(4), pp.885–897.

Lopilato, J., Bortner, S. & Beckwith, J., 1986. Mutations in a new chromosomal gene of Escherichia coli K-12, pcnB, reduce plasmid copy number of pBR322 and its derivatives. *Molecular & general genetics: MGG*, 205(2), pp.285–290.

Lovett, S.T., 2004. Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Molecular microbiology*, 52(5), pp.1243–1253.

Lovett, S.T. et al., 1994. Recombination between repeats in Escherichia coli by a recA-independent, proximity-sensitive mechanism. *Molecular & general genetics: MGG*, 245(3), pp.294–300.

Lovett, S.T. & Feschenko, V.V., 1996. Stabilization of diverged tandem repeats by mismatch repair: evidence for deletion formation via a misaligned replication intermediate. *Proceedings of the National Academy of Sciences of the United States of America*, 93(14), p.7120.

Low, K.O., Muhammad Mahadi, N. & Md. Illias, R., 2013. Optimisation of signal peptide for recombinant protein secretion in bacterial hosts. *Applied microbiology and biotechnology*, 97(9), pp.3811–3826.

Madar, D. et al., 2013. Promoter activity dynamics in the lag phase of Escherichia coli. *BMC systems biology*, 7, p.136.

Maddamsetti, R., Lenski, R.E. & Barrick, J.E., 2015. Adaptation, Clonal Interference, and Frequency-Dependent Interactions in a Long-Term Evolution Experiment with Escherichia coli. *Genetics*, 200(2), pp.619–631.

Manen, D., Goebel, T. & Caro, L., 1991. The par region of pSC101 affects plasmid copy number as well as stability. *Molecular microbiology*, 5(12), p.3087.

Marbach, A. & Bettenbrock, K., 2012. lac operon induction in Escherichia coli: Systematic comparison of IPTG and TMG induction and influence of the transacetylase LacA. *Journal of biotechnology*, 157(1), pp.82–88.

Marisch, K. et al., 2013. A Comparative Analysis of Industrial Escherichia coli K–12 and B Strains in High-Glucose Batch Cultivations on Process-, Transcriptome- and Proteome Level. *PloS one*, 8(8), p.e70516.

MarketsandMarkets, 2022. Recombinant Proteins Market. *MarketsandMarkets*. Available at: https://www.marketsandmarkets.com/Market-Reports/recombinant-proteins-market-70095015.html [Accessed January 16, 2024].

Marschall, L., Sagmeister, P. & Herwig, C., 2016. Tunable recombinant protein expression in E. coli: promoter systems and genetic constraints. *Applied microbiology and biotechnology*, 101(2), pp.501–512.

Martínez, J.L. et al., 2012. Pharmaceutical protein production by yeast: towards production of human blood proteins by microbial fermentation. *Current opinion in biotechnology*, 23(6), pp.965–971.

Massé, E., Vanderpool, C.K. & Gottesman, S., 2005. Effect of RyhB Small RNA on Global Iron Use in Escherichia coli. *Journal of bacteriology*, 187(20), p.6962.

Mazin, A.V. et al., 1991. Mechanisms of deletion formation in Escherichia coli plasmids. II. Deletions mediated by short direct repeats. *Molecular & general genetics: MGG*, 228(1-2). Available at: https://pubmed.ncbi.nlm.nih.gov/1679526/ [Accessed April 14, 2022].

McAllister, W.T. et al., 1981. Utilization of bacteriophage T7 late promoters in recombinant plasmids during infection. *Journal of molecular biology*, 153(3), pp.527–544.

Menacho-Melgar, R. et al., 2020. Scalable, two-stage, autoinduction of recombinant protein expression in E. coli utilizing phosphate depletion. *Biotechnology and bioengineering*, 117(9). Available at: https://pubmed.ncbi.nlm.nih.gov/32441815/ [Accessed September 4, 2023].

Merck KGaA, Enzymatic Assay of Chloramphenicol Acetyltransferase. *Enzyme Activity Assays, Merck*. Available at: https://www.sigmaaldrich.com/GB/en/technical-documents/protocol/protein-biology/enzyme-activity-assays/enzymatic-assay-of-chloramphenicol-acetyltransferase [Accessed August, 1 2023].

Millan, A.S. et al., 2015. Interactions between horizontally acquired genes create a fitness cost in Pseudomonas aeruginosa. *Nature Communications*, 6(1). Available at: http://dx.doi.org/10.1038/ncomms7845.

Million-Weaver, S. et al., 2012. Quantifying plasmid copy number to investigate plasmid dosage effects associated with directed protein evolution. *Methods in molecular biology* , 834, pp.33–48.

Minkley, E.G. & Pribnow, D., 1973. Transcription of the early region of bacteriophage T7: selective initiation with dinucleotides. *Journal of molecular biology*, 77(2), pp.255–277.

Mishra, R.K. & Chatterji, D., 1993. Mechanism of initiation of transcription by Escherichia coli RNA polymerase on supercoiled template. *Molecular microbiology*, 8(3), pp.507–515.

Mittl, P.R.E. & Schulz, G.E., 1994. Structure of glutathione reductase from escherichia coli at 1.86 Å resolution: Comparison with the enzyme from human erythrocytes. *Protein science: a publication of the Protein Society*, 3(5), pp.799–809.

Modi, R.I. & Adams, J., 1991. COEVOLUTION IN BACTERIAL-PLASMID POPULATIONS. *Evolution; international journal of organic evolution*, 45(3), pp.656–667.

Møller, T.S.B. et al., 2016. Relation between tetR and tetA expression in tetracycline resistant Escherichia coli. *BMC microbiology*, 16(1), pp.1–8.

Morrow, J.F. et al., 1974. Replication and Transcription of Eukaryotic DNA in Esherichia coli. *Proceedings of the National Academy of Sciences (PNAS)*, 71(5), pp.1743–1747.

Neubauer, P., Lin, H.Y. & Mathiszik, B., 2003. Metabolic load of recombinant protein production: Inhibition of cellular capacities for glucose uptake and respiration after induction of a heterologous gene in Escherichia coli. *Biotechnology and bioengineering*, 83(1), pp.53–64.

Niegemann, E., Schulz, A. & Bartsch, K., 1993. Molecular organization of the Escherichia coli gab cluster: nucleotide sequence of the structural genes gabD and gabP and expression of the GABA permease gene. *Archives of microbiology*, 160(6). Available at: https://pubmed.ncbi.nlm.nih.gov/8297211/ [Accessed April 27, 2024].

Nilsson, G. et al., 1984. Growth-rate dependent regulation of mRNA stability in Escherichia coli. *Nature*, 312(5989), pp.75–77.

Olivares-Hernández, R., Bordel, S. & Nielsen, J., 2011. Codon usage variability determines the correlation between proteome and transcriptome fold changes. *BMC systems biology*, 5(1), pp.1–9.

de Oliveira, J.D. et al., 2016. Genetic basis for hyper production of hyaluronic acid in natural and engineered microorganisms. *Microbial cell factories*, 15(1), pp.1–19.

Olsen, M.J. et al., 2003. High-Throughput FACS Method for Directed Evolution of Substrate Specificity. *Directed Enzyme Evolution*, pp.329–342.

Pandi, K. et al., 2020. Phosphate starvation controls lactose metabolism to produce recombinant protein in Escherichia coli. *Applied microbiology and biotechnology*, 104(22), pp.9707–9718.

Parvathy, S.T., Udayasuriyan, V. & Bhadana, V., 2021. Codon usage bias. *Molecular biology reports*, 49(1), pp.539–565.

Pasqua, M. et al., 2021. Modulation of OMV Production by the Lysis Module of the DLP12 Defective Prophage of Escherichia coli K12. *Microorganisms*, 9(2). Available at: https://pubmed.ncbi.nlm.nih.gov/33673345/ [Accessed April 28, 2024].

Pédelacq, J.-D. et al., 2006. Engineering and characterization of a superfolder green fluorescent protein. *Nature biotechnology*, 24(1), pp.79–88.

Peretti, S.W., Bailey, J.E. & Lee, J.J., 1989. Transcription from plasmid genes, macromolecular stability, and cell-specific productivity in Escherichia coli carrying copy number mutant plasmids. *Biotechnology and bioengineering*, 34(7), pp.902–908.

Peterman, N., Lavi-Itzkovitz, A. & Levine, E., 2014. Large-scale mapping of sequence-function relations in small regulatory RNAs reveals plasticity and modularity. *Nucleic acids research*, 42(19), p.12177.

Phue, J.N. et al., 2005. Glucose metabolism at high density growth of E. coli B and E. coli K: differences in metabolic pathways are responsible for efficient glucose utilization in E. coli B as determined by microarrays and Northern blot analyses. *Biotechnology and bioengineering*, 90(7). Available at: https://pubmed.ncbi.nlm.nih.gov/15806547/ [Accessed January 17, 2024].

Phue, J.N. et al., 2008. Modified Escherichia coli B (BL21), a superior producer of plasmid DNA compared with Escherichia coli K (DH5alpha). *Biotechnology and bioengineering*, 101(4). Available at: https://pubmed.ncbi.nlm.nih.gov/18814292/ [Accessed February 13, 2024].

Phue, J.N. & Shiloach, J., 2004. Transcription levels of key metabolic genes are the cause for different glucose utilization pathways in E. coli B (BL21) and E. coli K (JM109). *Journal of biotechnology*, 109(1-2). Available at: https://pubmed.ncbi.nlm.nih.gov/15063611/ [Accessed January 17, 2024].

Pletnev, P. et al., 2020. Comprehensive Functional Analysis of Escherichia coli Ribosomal RNA Methyltransferases. *Frontiers in genetics*, 11, p.472170.

Py, B. et al., 1996. A DEAD-box RNA helicase in the Escherichia coli RNA degradosome. *Nature*, 381(6578). Available at: https://pubmed.ncbi.nlm.nih.gov/8610017/ [Accessed April 28, 2024].

Qiao, J. et al., 2021. Construction of an Escherichia coli Strain Lacking Fimbriae by Deleting 64 Genes and Its Application for Efficient Production of Poly(3-Hydroxybutyrate) and l-Threonine. *Applied and environmental microbiology*, 87(12). Available at: https://pubmed.ncbi.nlm.nih.gov/33863704/ [Accessed April 28, 2024].

Quianzon, C.C. & Cheikh, I., 2012. History of insulin. *Journal of Community Hospital Internal Medicine Perspectives*, 2(2). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3714061/ [Accessed February 7, 2024].

Rahmen, N. et al., 2015. Exchange of single amino acids at different positions of a recombinant protein affects metabolic burden in Escherichia coli. *Microbial cell factories*, 14(1), pp.1–18.

Rajput, R., Sharma, R. & Gupta, R., 2011. Cloning and characterization of a thermostable detergent-compatible recombinant keratinase from Bacillus pumilus KS12. *Biotechnology and applied biochemistry*, 58(2), pp.109–118.

Rashid, M.H., 2022. Full-length recombinant antibodies from Escherichia coli: production, characterization, effector function (Fc) engineering, and clinical evaluation. *mAbs*. Available at: https://www.tandfonline.com/doi/abs/10.1080/19420862.2022.2111748 [Accessed February 7, 2024].

Ratzkin, B. & Carbon, J., 1977. Functional expression of cloned yeast DNA in Escherichia coli. *Proceedings of the National Academy of Sciences (PNAS)*, 74(2), pp.487–491.

Reyes-Lamothe, R. et al., 2014. High-copy bacterial plasmids diffuse in the nucleoid-free space, replicate stochastically and are randomly partitioned at cell division. *Nucleic acids research*, 42(2), pp.1042–1051.

Robertson, A.B. et al., 2006. MutL-catalyzed ATP hydrolysis is required at a post-UvrD loading step in methyl-directed mismatch repair. *The Journal of biological chemistry*, 281(29). Available at: https://pubmed.ncbi.nlm.nih.gov/16690604/ [Accessed April 29, 2024].

Roncero, C. & Casadaban, M.J., 1992. Genetic analysis of the genes involved in synthesis of the lipopolysaccharide core in Escherichia coli K-12: three operons in the rfa locus. *Journal of*

bacteriology, 174(10). Available at: https://pubmed.ncbi.nlm.nih.gov/1577693/ [Accessed April 28, 2024].

Rosano, G.L., Morales, E.S. & Ceccarelli, E.A., 2019. New tools for recombinant protein production in Escherichia coli: A 5-year update. *Protein science: a publication of the Protein Society*, 28(8), pp.1412–1422.

Rouches, M.V. et al., 2022. A plasmid system with tunable copy number. *Nature communications*, 13. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9263177/ [Accessed January 22, 2024].

Rowen, L. & Kornberg, A., 1978. Primase, the dnaG protein of Escherichia coli. An enzyme which starts DNA chains. *The Journal of biological chemistry*, 253(3). Available at: https://pubmed.ncbi.nlm.nih.gov/340457/ [Accessed April 28, 2024].

Sadler, J.R., Sasmor, H. & Betz, J.L., 1983. A perfectly symmetric lac operator binds the lac repressor very tightly. *Proceedings of the National Academy of Sciences of the United States of America*, 80(22), pp.6785–6789.

San Millan, A. et al., 2018. Integrative analysis of fitness and metabolic effects of plasmids in Pseudomonas aeruginosa PAO1. *The ISME journal*, 12(12), pp.3014–3024.

San Millan, A. & MacLean, R.C., 2017. Fitness Costs of Plasmids: a Limit to Plasmid Transmission. *Microbiology spectrum*, 5(5). Available at: http://dx.doi.org/10.1128/microbiolspec.MTBP-0016-2017.

Sereika, M. et al., 2022. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nature methods*, 19(7), pp.823–826.

Sergiev, P.V. et al., 2008. The ybiN gene of Escherichia coli encodes adenine-N6 methyltransferase specific for modification of A1618 of 23 S ribosomal RNA, a methylated residue located close to the ribosomal exit tunnel. *Journal of molecular biology*, 375(1). Available at: https://pubmed.ncbi.nlm.nih.gov/18021804/ [Accessed April 28, 2024].

Shaver, A.C. et al., 2002. Fitness Evolution and the Rise of Mutator Alleles in Experimental Escherichia coli Populations. *Genetics*, 162(2), pp.557–566.

Sheng, J., Ling, P. & Wang, F., 2015. Constructing a recombinant hyaluronic acid biosynthesis operon and producing food-grade hyaluronic acid in Lactococcus lactis. *Journal of industrial microbiology & biotechnology*, 42(2), pp.197–206.

Shimada, T. et al., 2019. Regulatory Role of PlaR (YiaJ) for Plant Utilization in Escherichia coli K-12. *Scientific reports*, 9(1). Available at: https://pubmed.ncbi.nlm.nih.gov/31892694/ [Accessed April 28, 2024].

Shukla, A.K. & Roy, K.B., 2006. Rec A-independent homologous recombination induced by a putative fold-back tetraplex DNA. *Biological chemistry*, 387(3). Available at: https://pubmed.ncbi.nlm.nih.gov/16542145/ [Accessed May 16, 2022].

Sieben, M. et al., 2016. Testing plasmid stability of Escherichia coli using the Continuously Operated Shaken BIOreactor System. *Biotechnology progress*, 32(6), pp.1418–1425.

Silvaa, S.A. e., Echeverrigaraya, S. & Gerhardt, G.J.L., 2011. BacPP: Bacterial promoter prediction—A tool for accurate sigma-factor specific assignment in enterobacteria. *Journal of theoretical biology*, 287, pp.92–99.

Simons, A. et al., 1984. Possible ideal lac operator: Escherichia coli lac operator-like sequences from eukaryotic genomes lack the central G X C pair. *Proceedings of the National Academy of Sciences of the United States of America*, 81(6), pp.1624–1628.

Sledjeski, D.D. & Gottesman, S., 1996. Osmotic shock induction of capsule synthesis in Escherichia coli K-12. *Journal of bacteriology*, 178(4), p.1204.

Smith, D.B. & Johnson, K.S., 1988. Single-step purification of polypeptides expressed in Escherichia coli as fusions with glutathione S-transferase. *Gene*, 67(1). Available at: https://pubmed.ncbi.nlm.nih.gov/3047011/ [Accessed February 12, 2024].

Smith, M.A. & Bidochka, M.J., 1998. Bacterial fitness and plasmid loss: the importance of culture conditions and plasmid size. *Canadian journal of microbiology*, 44(4), pp.351–355.

Soksawatmaekhin, W. et al., 2004. Excretion and uptake of cadaverine by CadB and its physiological functions in Escherichia coli. *Molecular microbiology*, 51(5). Available at: https://pubmed.ncbi.nlm.nih.gov/14982633/ [Accessed April 27, 2024].

del Solar, G. et al., 1998. Replication and Control of Circular Bacterial Plasmids. *Microbiology and molecular biology reviews: MMBR*. Available at: https://journals.asm.org/doi/10.1128/mmbr.62.2.434-464.1998 [Accessed January 22, 2024].

Song, Y. et al., 2015. Determination of single nucleotide variants in Escherichia coli DH5α by using short-read sequencing. *FEMS microbiology letters*, 362(11). Available at: https://academic.oup.com/femsle/article-pdf/362/11/fnv073/23924165/fnv073.pdf [Accessed November 18, 2021].

Son, M.S. & Taylor, R.K., 2021. Growth and Maintenance of Escherichia coli Laboratory Strains. *Current protocols*, 1(1), p.e20.

Sprouffske, K. et al., 2018. High mutation rates limit evolutionary adaptation in Escherichia coli. *PLoS genetics*, 14(4), p.e1007324.

Stalder, T. et al., 2017. Emerging patterns of plasmid-host coevolution that stabilize antibiotic resistance. *Scientific reports*, 7(1). Available at: https://pubmed.ncbi.nlm.nih.gov/28687759/ [Accessed February 13, 2024].

Steinhilper, R. et al., 2022. Structure of the membrane-bound formate hydrogenlyase complex from Escherichia coli. *Nature communications*, 13(1), pp.1–13.

Storz, G., Opdyke, J.A. & Zhang, A., 2004. Controlling mRNA stability and translation with small, noncoding RNAs. *Current opinion in microbiology*, 7(2), pp.140–144.

Struhl, K., Cameron, J.R. & Davis, R.W., 1976. Functional genetic expression of eukaryotic DNA in Escherichia coli. *Proceedings of the National Academy of Sciences (PNAS)*, 73(5), pp.1471–1475.

Studier, F.W. & Moffatt, B.A., 1986. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *Journal of molecular biology*, 189(1). Available at: https://pubmed.ncbi.nlm.nih.gov/3537305/ [Accessed February 13, 2024].

Studier, W.F. et al., 2009. Understanding the Differences between Genome Sequences of Escherichia coli B Strains REL606 and BL21(DE3) and Comparison of the E. coli B and K-12 Genomes. *Journal of molecular biology*, 394(4), pp.653–680.

Subbiah, M. et al., 2011. Selection Pressure Required for Long-Term Persistence of *bla* $_{CMY-2}$ -Positive IncA/C Plasmids. *Applied and Environmental Microbiology*, 77(13), pp.4486–4493. Available at: http://dx.doi.org/10.1128/aem.02788-10.

Summers, D.K. & Sherratt, D.J., 1984. Multimerization of high copy number plasmids causes instability: Cole 1 encodes a determinant essential for plasmid monomerization and stability. *Cell*, 36(4), pp.1097–1103.

Summers, D.K. & Sherratt, D.J., 1988. Resolution of ColE1 dimers requires a DNA sequence implicated in the three-dimensional organization of the cer site. *The EMBO journal*, 7(3), pp.851–858.

Tabor, S., 1990. Expression Using the T7 RNA Polymerase/Promoter System. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, 11(1), pp.16.2.1–16.2.11.

Tang, Z. et al., 2021. Fate of antibiotic resistance genes in industrial-scale rapid composting of pharmaceutical fermentation residue: The role implications of microbial community structure and mobile genetic elements. *Environmental pollution* , 291, p.118155.

Tian, D. et al., 2023. Cell Sorting-Directed Selection of Bacterial Cells in Bigger Sizes Analyzed by Imaging Flow Cytometry during Experimental Evolution. *International journal of molecular sciences*, 24(4), p.3243.

Tolmasky, M. & Alonso, J.C., 2015. Mechanisms of Theta Plasmid Replication. *Microbiology Spectrum*. Available at: https://journals.asm.org/doi/10.1128/microbiolspec.plas-0029-2014 [Accessed January 22, 2024].

Tracy, B.P., Gaida, S.M. & Papoutsakis, E.T., 2010. Flow cytometry for bacteria: enabling metabolic engineering, synthetic biology and the elucidation of complex phenotypes. *Current opinion in biotechnology*, 21(1), pp.85–99.

Tripathi, L., Zhang, Y. & Lin, Z., 2014. Bacterial Sigma Factors as Targets for Engineered or Synthetic Transcriptional Control. *Frontiers in Bioengineering and Biotechnology*, 0. Available at: http://dx.doi.org/10.3389/fbioe.2014.00033 [Accessed April 25, 2022].

Tuggle, C.K. & Fuchs, J.A., 1985. Glutathione reductase is not required for maintenance of reduced glutathione in Escherichia coli K-12. *Journal of bacteriology*, 162(1). Available at: https://pubmed.ncbi.nlm.nih.gov/3884598/ [Accessed April 27, 2024].

Tyler, J. & Sherratt, D.J., 1975. Synthesis of E colicins in Escherichia coli. *Molecular & general genetics: MGG*, 140(4), pp.349–353.

Uhlin, B.E. & Nordström, K., 1977. R plasmid gene dosage effects in Escherichia coli K-12: copy mutants of the R plasmic R1drd-19. *Plasmid*, 1(1). Available at: https://pubmed.ncbi.nlm.nih.gov/375268/ [Accessed February 12, 2024].

Vapnek, D. et al., 1977. Expression in Escherichia coli K-12 of the structural gene for catabolic dehydroquinase of Neurospora crassa. *Proceedings of the National Academy of Sciences of the United States of America*, 74(8), pp.3508–3512.

Vyas, V.V., Gupta, S. & Sharma, P., 1994. Stability of a recombinant shuttle plasmid in Bacillus subtilis and Escherichia coli. *Enzyme and microbial technology*, 16(3), pp.240–246.

Wacker, M. et al., 2002. N-linked glycosylation in Campylobacter jejuni and its functional transfer into E. coli. *Science* , 298(5599). Available at: https://pubmed.ncbi.nlm.nih.gov/12459590/ [Accessed February 7, 2024].

Wang, B. et al., 2020. Risk of penicillin fermentation dreg: Increase of antibiotic resistance genes after soil discharge. *Environmental pollution* , 259, p.113956.

Wang, P. et al., 2010. Robust growth of Escherichia coli. *Current biology: CB*, 20(12). Available at: https://pubmed.ncbi.nlm.nih.gov/20537537/ [Accessed January 24, 2024].

Wang, Y., Penkul, P. & Milstein, J.N., 2016. Quantitative Localization Microscopy Reveals a Novel Organization of a High-Copy Number Plasmid. *Biophysical journal*, 111(3), pp.467–479.

Watt, V.M. et al., 1985. Homology requirements for recombination in Escherichia coli. *Proceedings of the National Academy of Sciences*, 82(14), pp.4768–4772. Available at: http://dx.doi.org/10.1073/pnas.82.14.4768.

Whitfield, C., Amor, P.A. & Köplin, R., 1997. Modulation of the surface architecture of gram-negative bacteria by the action of surface polymer:lipid A-core ligase and by determinants of polymer chain length. *Molecular microbiology*, 23(4). Available at: https://pubmed.ncbi.nlm.nih.gov/9157235/ [Accessed April 28, 2024].

Wick, R.R. & Holt, K.E., 2022. Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLoS computational biology*, 18(1). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8812927/ [Accessed April 26, 2024].

Wick, R.R., Judd, L.M. & Holt, K.E., 2023. Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing. *PLoS computational biology*, 19(3), p.e1010905.

Wielgoss, S. et al., 2011. Mutation Rate Inferred From Synonymous Substitutions in a Long-Term Evolution Experiment With Escherichia coli. *G3 Genes|Genomes|Genetics*, 1(3), pp.183–186.

Xia, X.X. et al., 2008. Comparison of the extracellular proteomes of Escherichia coli B and K-12 strains during high cell density cultivation. *Proteomics*, 8(10). Available at: https://pubmed.ncbi.nlm.nih.gov/18425732/ [Accessed January 17, 2024].

Xu, J. et al., 2012. Galactose can be an inducer for production of therapeutic proteins by auto-induction using E. coli BL21 strains. *Protein expression and purification*, 83(1). Available at: https://pubmed.ncbi.nlm.nih.gov/22425658/ [Accessed February 12, 2024].

Yang, G. & Withers, S.G., 2009. Ultrahigh-Throughput FACS-Based Screening for Directed Enzyme Evolution. *Chembiochem: a European journal of chemical biology*, 10(17), pp.2704–2715.

Yang, Y. & Sha, M., A Beginner's Guide to Bioprocess Modes – Batch, FedBatch, and Continuous Fermentation. Available at: https://www.eppendorf.com/product-media/doc/en/763594/Fermentors-Bioreactors_Application-Note_408_BioBLU-f-Single-Vessel_A-Beginner%E2%80%99s-Guide-Bioprocess-Modes-Batch_Fed-Batch-Continuous-Fermentation.pdf [Accessed February 13, 2024].

Yano, H. et al., 2016. Evolved plasmid-host interactions reduce plasmid interference cost. *Molecular microbiology*, 101(5), pp.743–756.

Yim, S. et al., 2001. High-level secretory production of human granulocyte-colony stimulating factor by fed-batch culture of recombinant Escherichia coli. *Bioprocess and biosystems engineering*, 24(4), pp.249–254.

Yin, J. et al., 2007. Select what you need: a comparative evaluation of the advantages and limitations of frequently used expression systems for foreign genes. *Journal of biotechnology*, 127(3). Available at: https://pubmed.ncbi.nlm.nih.gov/16959350/ [Accessed January 16, 2024].

Yoon, S.H. et al., 2003. Combined transcriptome and proteome analysis of Escherichia coli during high cell density culture. *Biotechnology and bioengineering*, 81(7), pp.753–767.

Yoon, S.H. et al., 2012. Comparative multi-omics systems analysis of Escherichia coli strains B and K-12. *Genome biology*, 13(5), pp.1–13.

Yoshida, M. et al., 2014. Directed evolution of cell size in Escherichia coli. *BMC evolutionary biology*, 14(1), pp.1–12.

Yurtsev, E.A. et al., 2013. Bacterial cheating drives the population dynamics of cooperative antibiotic resistance plasmids. *Molecular systems biology*, 9, p.683.

Zahavi, D. & Weiner, L., 2020. Monoclonal Antibodies in Cancer Therapy. *Antibodies*, 9(3), p.34.

Zhong, C., Wei, P. & Zhang, Y.P., 2017. Enhancing functional expression of codon-optimized heterologous enzymes in Escherichia coli BL21(DE3) by selective introduction of synonymous rare codons. *Biotechnology and bioengineering*, 114(5). Available at: https://pubmed.ncbi.nlm.nih.gov/27943233/ [Accessed February 12, 2024].

Zielenkiewicz, U. & Cegłowski, P., 2001. Mechanisms of plasmid stable maintenance with special focus on plasmid addiction systems. *Acta biochimica Polonica*, 48(4), pp.1003–1023.

Zuo, Z. & Stormo, G.D., 2014. High-Resolution Specificity from DNA Sequencing Highlights Alternative Modes of Lac Repressor Binding. *Genetics*, 198(3), pp.1329–1343. Available at: http://dx.doi.org/10.1534/genetics.114.170100.

8. Supplementary materials

Population number	Pre-exposure	Operator	Recombined promoter	Background
1	no	original	no	K
2	no	original	no	K
3	no	original	no	K
4	no	original	no	K
5	no	original	no	K
6	no	original	no	K
7	no	original	no	K
8	no	original	no	K
9	no	original	no	K
10	no	original	no	K
11	no	original	no	K
12	no	original	no	K
13	no	original	no	K
14	no	original	no	K
15	no	original	no	K
16	no	original	no	K
17	no	original	no	K

18	no	original	no	K
19	no	original	no	K
20	no	original	no	K
21*	no	original	yes	K
22*	no	original	no	K
23*	no	original	no	K
1	no	alternative	no	K
2	no	alternative	no	K
3	no	alternative	no	K
4	no	alternative	no	K
5	no	alternative	no	K
6	no	alternative	no	K
7	no	alternative	no	K
8	no	alternative	no	K
9	no	alternative	no	K
10	no	alternative	no	K
11	no	alternative	no	K
12	no	alternative	no	K
13	no	alternative	no	K
14	no	alternative	no	K

15	no	alternative	no	K
16	no	alternative	no	K
17	no	alternative	no	K
18	no	alternative	no	K
19	no	alternative	no	K
20	no	alternative	no	K
21*	no	alternative	no	K
22*	no	alternative	no	K
23*	no	alternative	no	K
1	no	original	no	B
2	no	original	no	B
3	no	original	no	B
4	no	original	no	B
5	no	original	no	B
6	no	original	no	B
7	no	original	no	B
8	no	original	no	B
9	no	original	no	B
10	no	original	no	B
11	no	original	no	B

12	no	original	no	B
13	no	original	no	B
14	no	original	no	B
15	no	original	no	B
16	no	original	no	B
17	no	original	no	B
18	no	original	no	B
19	no	original	no	B
20	no	original	no	B
21*	no	original	no	B
22*	no	original	no	B
23*	no	original	no	B
1	no	alternative	no	B
2	no	alternative	no	B
3	no	alternative	no	B
4	no	alternative	no	B
5	no	alternative	no	B
6	no	alternative	no	B
7	no	alternative	no	B
8	no	alternative	no	B

9	no	alternative	no	B
10	no	alternative	no	B
11	no	alternative	no	B
12	no	alternative	no	B
13	no	alternative	no	B
14	no	alternative	no	B
15	no	alternative	no	B
16	no	alternative	no	B
17	no	alternative	no	B
18	no	alternative	no	B
19	no	alternative	no	B
20	no	alternative	no	B
21*	no	alternative	no	B
22*	no	alternative	no	B
23*	no	alternative	no	B
1	yes	original	yes	K
2	yes	original	no	K
3	yes	original	no	K
4	yes	original	no	K
5	yes	original	yes	K

6	yes	original	no	K
7	yes	original	yes	K
8	yes	original	yes	K
9	yes	original	yes	K
10	yes	original	yes	K
11	yes	original	yes	K
12	yes	original	no	K
13	yes	original	yes	K
14	yes	original	yes	K
15	yes	original	yes	K
16	yes	original	no	K
17	yes	original	yes	K
18	yes	original	yes	K
19	yes	original	yes	K
20	yes	original	yes	K
21*	yes	original	yes	K
22*	yes	original	yes	K
23*	yes	original	no	K
1	yes	alternative	no	K
2	yes	alternative	no	K

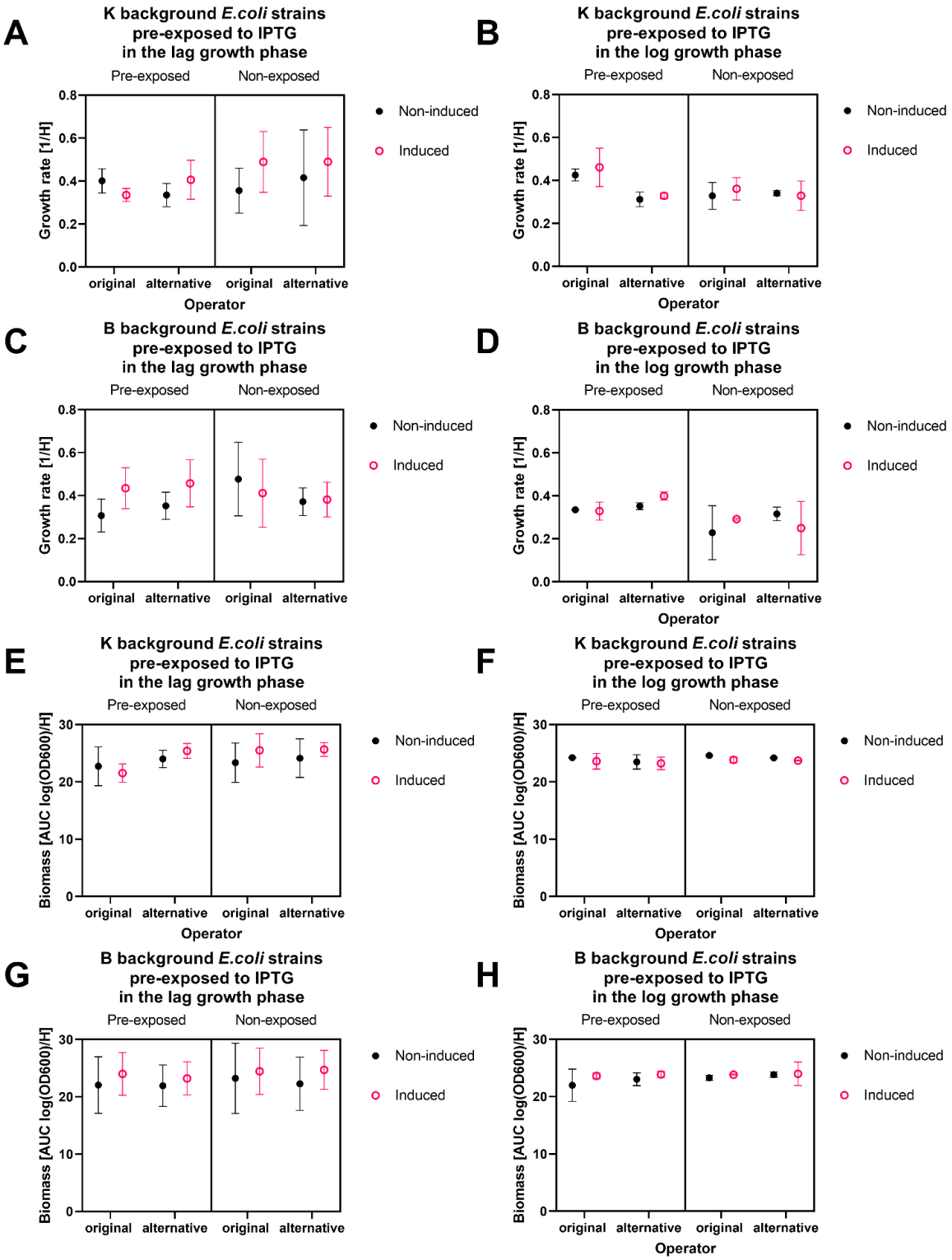
3	yes	alternative	no	K
4	yes	alternative	no	K
5	yes	alternative	no	K
6	yes	alternative	no	K
7	yes	alternative	no	K
8	yes	alternative	no	K
9	yes	alternative	no	K
10	yes	alternative	no	K
11	yes	alternative	no	K
12	yes	alternative	no	K
13	yes	alternative	no	K
14	yes	alternative	no	K
15	yes	alternative	no	K
16	yes	alternative	no	K
17	yes	alternative	no	K
18	yes	alternative	no	K
19	yes	alternative	no	K
20	yes	alternative	no	K
21*	yes	alternative	no	K
22*	yes	alternative	yes	K

23*	yes	alternative	no	K
1	yes	original	yes	B
2	yes	original	yes	B
3	yes	original	yes	B
4	yes	original	yes	B
5	yes	original	yes	B
6	yes	original	yes	B
7	yes	original	yes	B
8	yes	original	yes	B
9	yes	original	yes	B
10	yes	original	yes	B
11	yes	original	yes	B
12	yes	original	yes	B
13	yes	original	yes	B
14	yes	original	yes	B
15	yes	original	yes	B
16	yes	original	yes	B
17	yes	original	yes	B
18	yes	original	yes	B
19	yes	original	yes	B

20	yes	original	no	B
21*	yes	original	no	B
22*	yes	original	no	B
23*	yes	original	no	B
1	yes	alternative	no	B
2	yes	alternative	no	B
3	yes	alternative	no	B
4	yes	alternative	no	B
5	yes	alternative	no	B
6	yes	alternative	no	B
7	yes	alternative	no	B
8	yes	alternative	no	B
9	yes	alternative	no	B
10	yes	alternative	no	B
11	yes	alternative	no	B
12	yes	alternative	no	B
13	yes	alternative	no	B
14	yes	alternative	no	B
15	yes	alternative	no	B
16	yes	alternative	no	B

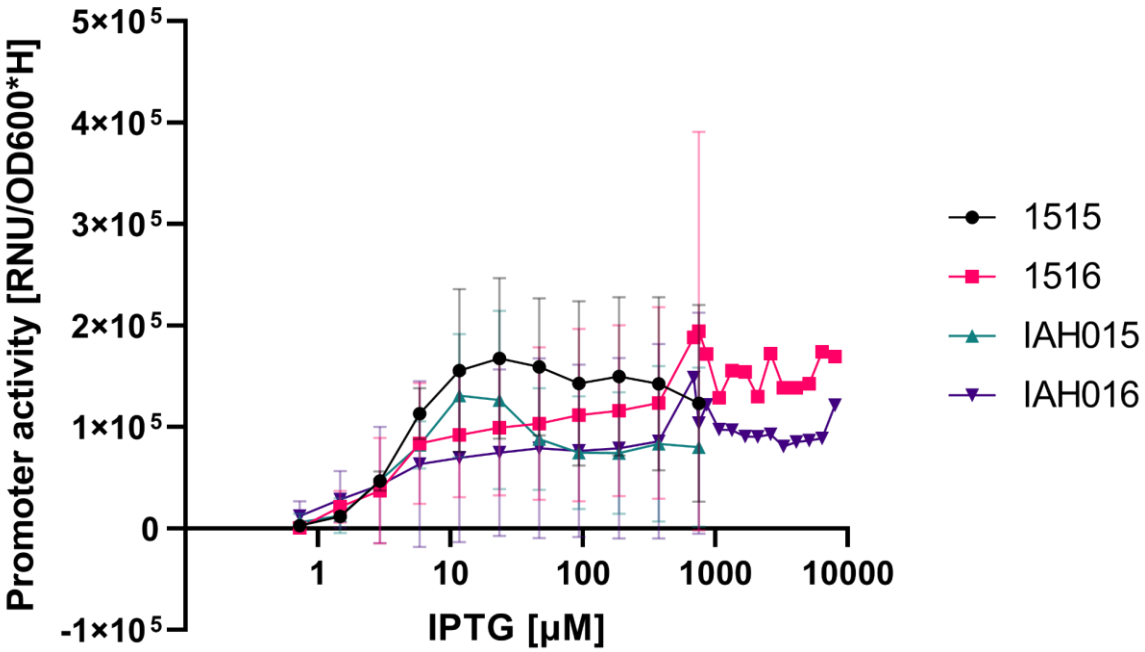
17	yes	alternative	no	B
18	yes	alternative	no	B
19	yes	alternative	no	B
20	yes	alternative	no	B
21*	yes	alternative	no	B
22*	yes	alternative	no	B
23*	yes	alternative	no	B

Supplementary table S1. The multivariable table was constructed after counting the recombination events from the gels presented in Fig. 5. It was then analysed using multiple logistic regression in GraphPad Prism 9.0.0 Asterisks indicate replicates included from an earlier experiment of a similar design.

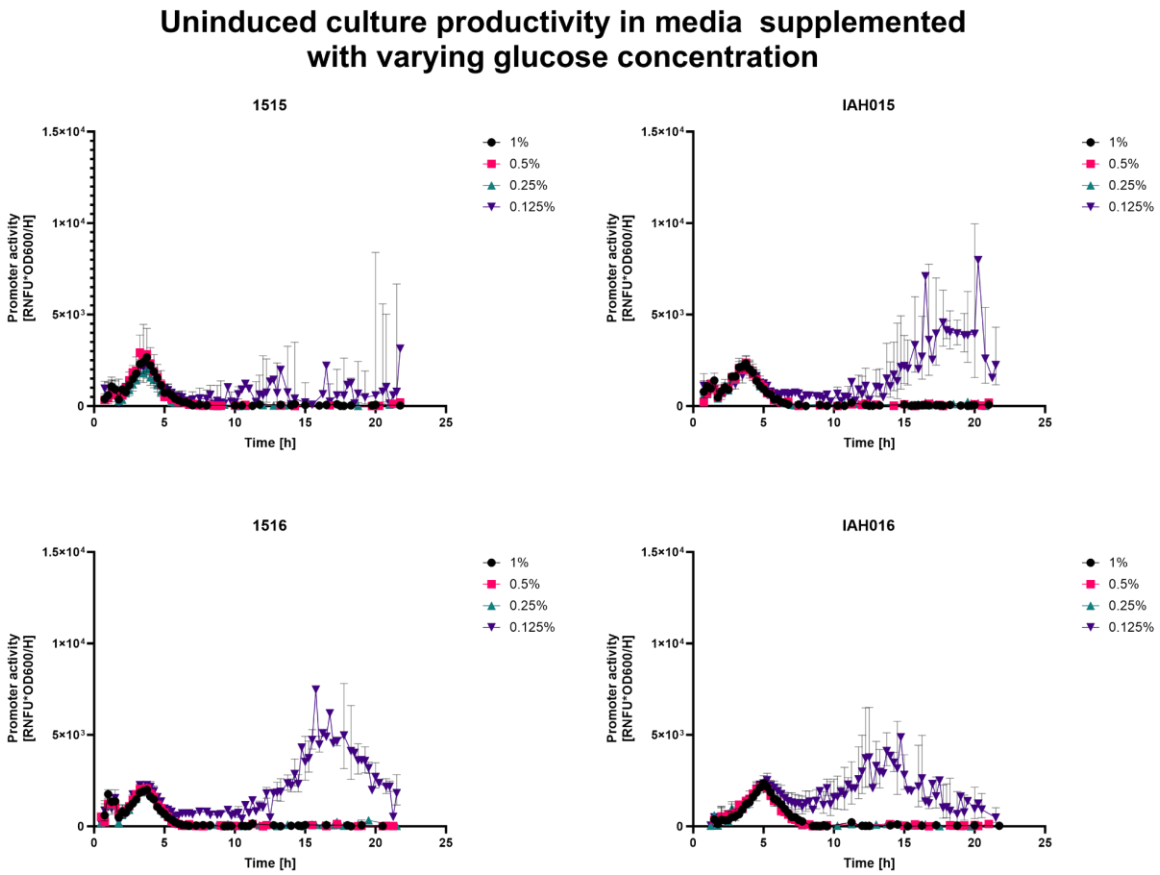


Supplementary Figure S2. Comparison of the growth rates (A, B, C, D) and biomass accumulation (E, F, G, H) in various E.coli strains carrying either pAVE011 (original) or pIAH011 (alternative) plasmid. The growth rates (A, B, C, D) and total biomass accumulation in 24h of growth (E, F, G, H) are compared. The

strains were pre-exposed to 0.5mM IPTG either in the lag (A, C, E, G) or exponential growth phase (B, D, F, H). K background E. coli strains (1515 and IAH015) are compared in A, B, E, and F, while B background (1516 and IAH016) are compared in C, D, G, and H.



Supplementary Figure S3. Total promoter activity in the 24h post-induction calculated as AUC of the promoter activity/time curves across a range of inducer concentrations. Each plotted point was calculated as a mean of 3 replicates. X axis has a logarithmic scale [log(10)], while Y axis has a linear scale.



Supplementary figure S4 Leaky (uninduced) culture productivity of various pAVEway strains in vLB supplemented with varying amounts of glucose. Each panel represents one strain. The percentages are the final glucose concentration in the media. Each datapoint is a geometric mean of three replicates and error bars are showing geometric SD.

Averaging method:	Used when:
Arithmetic mean	Datapoints are normally distributed
Geometric mean	Logarithms of datapoints are normally distributed
Harmonic mean	Reciprocals of datapoints are normally distributed
Quadratic mean	Datapoints squared are normally distributed
Median	Datapoints do not follow normal distribution patterns regardless of transformation applied

Supplementary Table S5 Averaging methods used in the study and the rationale used in their application. The rules were based on the “Intuitive Biostatistics” book and the online GraphPad Prism statistics guide.

Plate row	Plate column	Expected genotype	Population number	Clone number	Ancestral or evolved	Confirmed genotype
A	1	BL21 no plasmid	1	N/A	Ancestral	W3110 no plasmid
A	2	BL21 no plasmid	2	N/A	Ancestral	W3110 no plasmid
A	3	BL21 no plasmid	3	N/A	Ancestral	W3110 no plasmid
A	4	BL21 no plasmid	4	N/A	Ancestral	W3110 no plasmid
A	5	BL21 no plasmid	5	N/A	Ancestral	W3110 no plasmid
A	6	BL21 no plasmid	6	N/A	Ancestral	W3110 no plasmid
A	7	BL21 no plasmid	1	N/A	Evolved	W3110 no plasmid

A	8	BL21 no plasmid	2	N/A	Evolved	W3110 no plasmid
A	9	BL21 no plasmid	3	N/A	Evolved	W3110 no plasmid
A	10	BL21 no plasmid	4	N/A	Evolved	W3110 no plasmid
A	11	BL21 no plasmid	5	N/A	Evolved	BL21 no plasmid
A	12	BL21 no plasmid	6	N/A	Evolved	BL21 no plasmid
B	1	W3110 no plasmid	1	N/A	Ancestral	W3110 no plasmid
B	2	W3110 no plasmid	2	N/A	Ancestral	W3110 no plasmid
B	3	W3110 no plasmid	3	N/A	Ancestral	W3110 no plasmid
B	4	W3110 no plasmid	4	N/A	Ancestral	W3110 no plasmid
B	5	W3110 no plasmid	5	N/A	Ancestral	W3110 no plasmid
B	6	W3110 no plasmid	6	N/A	Ancestral	W3110 no plasmid
B	7	W3110 no plasmid	1	N/A	Evolved	BL21 no plasmid
B	8	W3110 no plasmid	2	N/A	Evolved	BL21 no plasmid
B	9	W3110 no plasmid	3	N/A	Evolved	W3110 no plasmid
B	10	W3110 no plasmid	4	N/A	Evolved	W3110 no plasmid
B	11	W3110 no plasmid	5	N/A	Evolved	BL21 no plasmid

B	12	W3110 no plasmid	6	N/A	Evolved	W3110 no plasmid
C	11	BL21 empty plasmid	1	N/A	Ancestral	W3110 empty plasmid
C	12	BL21 empty plasmid	2	N/A	Ancestral	W3110 empty plasmid
D	1	BL21 empty plasmid	3	N/A	Ancestral	W3110 empty plasmid
D	2	BL21 empty plasmid	4	N/A	Ancestral	W3110 empty plasmid
D	3	BL21 empty plasmid	5	N/A	Ancestral	W3110 empty plasmid
D	4	BL21 empty plasmid	6	N/A	Ancestral	W3110 empty plasmid
D	5	BL21 empty plasmid	1	N/A	Evolved	W3110 empty plasmid
D	6	BL21 empty plasmid	2	N/A	Evolved	W3110 empty plasmid
D	7	BL21 empty plasmid	3	N/A	Evolved	W3110 empty plasmid
D	8	BL21 empty plasmid	4	N/A	Evolved	W3110 empty plasmid

D	9	BL21 empty plasmid	5	N/A	Evolved	W3110 empty plasmid
D	10	BL21 empty plasmid	6	N/A	Evolved	W3110 empty plasmid
D	11	W3110 empty plasmid	1	N/A	Ancestral	W3110 empty plasmid
D	12	W3110 empty plasmid	2	N/A	Ancestral	W3110 empty plasmid
E	1	W3110 empty plasmid	3	N/A	Ancestral	W3110 empty plasmid
E	2	W3110 empty plasmid	4	N/A	Ancestral	W3110 empty plasmid
E	3	W3110 empty plasmid	5	N/A	Ancestral	W3110 empty plasmid
E	4	W3110 empty plasmid	6	N/A	Ancestral	W3110 empty plasmid
E	5	W3110 empty plasmid	1	N/A	Evolved	W3110 empty plasmid
E	6	W3110 empty plasmid	2	N/A	Evolved	W3110 empty plasmid
E	7	W3110 empty plasmid	3	N/A	Evolved	W3110 empty plasmid

E	8	W3110 empty plasmid	4	N/A	Evolved	W3110 empty plasmid
E	9	W3110 empty plasmid	5	N/A	Evolved	W3110 empty plasmid
E	10	W3110 empty plasmid	6	N/A	Evolved	W3110 empty plasmid
E	11	1516	1	10	Ancestral	IAH015
E	12	1516	2	4	Ancestral	IAH015
F	1	1516	3	11	Ancestral	IAH015
F	2	1516	4	7	Ancestral	IAH015
F	3	1516	5	8	Ancestral	IAH015
F	4	1516	6	2	Ancestral	IAH015
F	5	IAH016	1	11	Ancestral	IAH015
F	6	IAH016	2	8	Ancestral	IAH015
F	7	IAH016	3	9	Ancestral	IAH015
F	8	IAH016	4	11	Ancestral	IAH015
F	9	IAH016	5	6	Ancestral	IAH015
F	10	IAH016	6	10	Ancestral	IAH015
F	11	1515	1	6	Ancestral	1515
F	12	1515	2	12	Ancestral	1515
G	1	1515	3	9	Ancestral	1515
G	2	1515	4	8	Ancestral	1515
G	3	1515	5	11	Ancestral	1515
G	4	1515	6	7	Ancestral	1515
G	5	IAH015	1	1	Ancestral	IAH015
G	6	IAH015	2	12	Ancestral	IAH015
G	7	IAH015	3	3	Ancestral	IAH015
G	8	IAH015	4	2	Ancestral	IAH015
G	9	IAH015	5	10	Ancestral	IAH015
G	10	IAH015	6	7	Ancestral	IAH015

Supplementary Table S6. Populations and clones chosen to be sequenced on the "control" plate. This includes ancestral populations as well as evolved populations of the control (plasmid-free and empty plasmid-carrying) lineages.

Plate row	Plate column	Expected genotype	Population number	Clone number	Phenotype group	Confirmed genotype
A	1	IAH015	3	3	PULU	IAH015
A	2	IAH015	3	5	PULU	IAH015
A	3	IAH015	3	6	PULU	IAH015
A	4	IAH015	3	7	PULU	IAH015
A	5	IAH015	3	9	PULU	IAH015
A	6	IAH015	3	10	PULU	IAH015
A	7	IAH015	3	11	PULU	IAH015
A	8	IAH016	1	4	PULU	IAH015
A	9	1516	5	10	PULU	IAH015
A	10	IAH015	3	1	PULS	IAH015
A	11	IAH015	3	2	PULS	IAH015
A	12	IAH015	3	4	PULS	IAH015
B	1	IAH015	3	8	PULS	IAH015
B	2	IAH015	3	12	PULS	IAH015
B	3	IAH015	4	1	PULS	IAH015
B	4	IAH015	4	8	PULS	IAH015
B	5	IAH015	4	10	PULS	IAH015
B	6	IAH015	4	11	PULS	IAH015
B	7	IAH015	5	7	PULS	IAH015
B	8	IAH015	5	10	PULS	IAH015
B	9	IAH016	6	1	PULD	IAH015
B	10	IAH016	6	4	PULD	IAH015
B	11	IAH016	6	10	PULD	IAH015
B	12	IAH016	6	11	PULD	IAH015
C	1	IAH016	6	6	PULD	IAH015
C	2	1516	1	11	PULD	IAH015
C	3	1516	1	8	PULD	IAH015
C	4	1516	1	9	PULD	IAH015
C	5	1516	2	4	PULD	IAH015
C	6	1516	2	10	PULD	IAH015
C	7	1516	2	5	PULD	IAH015
C	8	1516	2	12	PULD	IAH015
C	9	1516	3	3	PULD	IAH015

C	10	1516	3	10	PULD	IAH015
C	11	1516	3	11	PULD	IAH015
C	12	1516	4	7	PULD	IAH015
D	1	1516	4	11	PULD	IAH015
D	2	1516	4	6	PULD	IAH015
D	3	1516	5	8	PULD	IAH015
D	4	1516	5	12	PULD	IAH015
D	5	1516	5	7	PULD	IAH015
D	6	1516	6	3	PULD	IAH015
D	7	1516	6	5	PULD	IAH015
D	8	1516	6	7	PULD	IAH015
D	9	1516	6	4	PULD	IAH015
D	10	1515	1	6	PSLD	1515
D	11	1515	2	11	PSLD	1515
D	12	1515	3	2	PSLD	1515
E	1	1515	3	10	PSLD	1515
E	2	1515	4	2	PSLD	1515
E	3	1515	4	4	PSLD	1515
E	4	1515	5	4	PSLD	1515
E	5	1515	6	10	PSLD	1515
E	6	1515	6	11	PSLD	1515
E	7	IAH016	1	8	PSLD	IAH015
E	8	IAH016	2	9	PSLD	IAH015
E	9	IAH016	2	10	PSLD	IAH015
E	10	IAH016	3	9	PSLD	IAH015
E	11	IAH016	3	4	PSLD	IAH015
E	12	IAH016	4	11	PSLD	IAH015
F	1	IAH016	5	7	PSLD	IAH015
F	2	IAH016	6	3	PSLD	IAH015
F	3	IAH016	6	2	PSLD	IAH015
F	4	1516	1	4	PSLD	IAH015
F	5	1516	1	2	PSLD	IAH015
F	6	1516	3	1	PSLD	IAH015
F	7	1516	3	12	PSLD	IAH015
F	8	1516	3	6	PSLD	IAH015
F	9	1516	5	3	PSLD	IAH015
F	10	1516	6	11	PSLD	IAH015
F	11	1516	6	2	PSLD	IAH015
F	12	1516	6	12	PSLD	IAH015
G	1	1515	1	12	PDLD	1515
G	2	1515	2	1	PDLD	1515
G	3	1515	3	12	PDLD	1515

G	4	1515	4	1	PDLD	1515
G	5	1515	5	7	PDLD	1515
G	6	1515	5	3	PDLD	1515
G	7	1515	6	9	PDLD	1515
G	8	1515	6	3	PDLD	1515
G	9	IAH015	1	4	PDLD	IAH015
G	10	IAH015	1	6	PDLD	IAH015
G	11	IAH015	4	3	PDLD	IAH015
G	12	IAH015	4	2	PDLD	IAH015
H	1	IAH015	5	9	PDLD	IAH015
H	2	IAH015	5	3	PDLD	IAH015
H	3	IAH015	6	5	PDLD	IAH015
H	4	IAH015	6	8	PDLD	IAH015
H	5	IAH016	2	5	PDLD	IAH015
H	6	IAH016	2	6	PDLD	IAH015
H	7	IAH016	4	3	PDLD	IAH015
H	8	IAH016	4	7	PDLD	IAH015
H	9	IAH016	4	10	PDLD	IAH015
H	10	IAH016	5	5	PDLD	IAH015
H	11	IAH016	6	7	PDLD	IAH015
H	12	IAH016	6	5	PDLD	IAH015

Supplementary Table S7. Populations and clones chosen to be sequenced on the "evolved" plate. This includes all stable evolved clones chosen for sequencing.