A functional analysis of random coding sequences in Escherichia coli

Dissertation

in fulfilment of the requirements for the degree

Doctor rerum naturalium (Dr. rer. nat.)

of the Faculty of Mathematics and Natural Sciences

at Kiel University

submitted by

Devika Bhave

Department of Evolutionary Genetics

Max Planck Institute for Evolutionary Biology

Plön, June 2020

First examiner: Prof. Dr. Diethard Tautz

Second examiner: Prof. Dr. Ruth Schmitz-Streit

Additional examiner: Prof. Dr. Paul Rainey

Chairperson: Prof. Dr. Regina Scherließ

Date of the oral examination: 11th September 2020

In the loving memory of my grandmother, Shobhana Prabhakar Joshi (1941-2019)

Zusammenfassung

Die Anpassung von Organismen an sich ständig verändernde Umgebungen beinhaltet die Generierung genetischer Neuheit in ihren Genomen durch Mechanismen wie *De novo*-Genevolution, Duplikation, Fusion, lateralen Gentransfer usw. *De novo*-Genevolution ist ein Mechanismus, bei dem sich neue Genfunktionen aus zuvor nicht-kodierenden Sequenzen entwickeln können, bei denen es sich im Wesentlichen um zufällige Abschnitte von Nukleotiden handelt. Mehrere Studien haben die Rolle solcher Zufallssequenzen als Ausgangspunkt für evolutionäre Innovation untersucht. Dazu gehörte eine systematische Studie, bei der eine Bibliothek von zufälligen Codierungssequenzen in *E. coli* exprimiert und das differentielle Wachstum gemessen wurde, um die Fitnesseffekte einzelner Sequenzen zu beurteilen. Jede zufällige Sequenz aus der Bibliothek wurde auf der Grundlage der Veränderung ihrer Häufigkeit in der Population im Laufe der Zeit in negativ, positiv oder neutral kategorisiert. In dieser Arbeit analysiere ich die Auswirkungen einzelner Klone, die von diesem Screen abgeleitet wurden.

Um die Auswirkungen von Zufallssequenzen auf die Fitness des Wirts zu untersuchen, klonte ich repräsentative Varianten aus jeder Kategorie mit Hilfe eines Multikopie-Plasmidvektors in *E. coli*-Stämme. Im ersten Teil der Arbeit zeige ich, dass die Expression negativer Zufallspeptide einen Fitness-Nachteil (schädlich) in *E. coli* verursacht, gefolgt von einer Wachstumserholung. Nach weiteren Untersuchungen fand ich, dass diese Peptide unmittelbar nach der Expression eine Stressantwort im Wirt hervorrufen. Der hochgradig schädliche Phänotyp kann so im Wirt kompensiert werden. Darüber hinaus konnte ich Phänotyp-Suppressor-Klone isolieren. Die Resequenzierung der Suppressoren zusammen mit jedem der Ausgangsklone half bei der Identifizierung von Interaktionspartnern für die schädlichen Peptide. Im zweiten Teil zeige ich zwei Mechanismen, die der Wirt nutzt, um sich an eine schädliche Peptidexpression anzupassen: (a) Kontrolle der Plasmidkopienzahl durch Inaktivierung des pcnB-Gens und (b) Expressionskontrolle durch Inaktivierung der LacI-Inducer-Bindungsdomäne. Im dritten Teil der Arbeit zeige ich, dass die positiven

Zufallspeptide unter Stressbedingungen, z.B. bei erhöhter Temperatur, einen Fitnessvorteil verschaffen. Abschließend zeige ich, dass Zufallssequenzen in der Tat die Fitness des Wirtes beeinflussen, möglicherweise durch das gezielte Aktivieren bestimmter Gene oder Proteine. Diese Studie liefert experimentelle Hinweise darauf, wie Zufallssequenzen als Treiber der *de novo*-Genevolution dienen könnten.

(This abstract was kindly translated into German by Prof. Dr. Diethard Tautz)

Abstract

Adaptation of organisms to continuously changing environments includes the generation of genic novelty in their genomes through mechanisms such as *de novo* gene evolution, duplication, fusion, lateral gene transfer, etc. *De novo* gene evolution is a mechanism, wherein new gene functions can evolve from previously non-coding sequences, which are essentially random stretches of nucleotides. Several studies have explored the role of such random sequences as templates for evolutionary innovation. This included a systematic study, where a library of random coding sequences was expressed in *Escherichia coli* and differential growth was measured to assess fitness effects of individual sequences. Each random sequence from the library was categorized into negative, positive or neutral based on its change in abundance in the population across time. In this thesis, I analyse the effects of individual clones derived from this screen.

In order to study effects of random sequences on the fitness of the host, I cloned representative variants from each category into *E. coli* strains using a multicopy plasmid vector. In the first part of the thesis, I demonstrate that expression of negative random peptides confers a fitness disadvantage (deleterious) in *E. coli*, followed by a growth recovery. Upon further investigation, I find that these peptides can elicit a stress response in the host instantaneously upon expression. The highly deleterious phenotype can thus be compensated in the host. In addition, I was able to isolate suppressor-of-phenotype clones. Re-sequencing of the suppressors together with each of the ancestor clones helped identify interaction partners for the deleterious peptides. In the second part, I show two mechanisms that the host uses to adapt to deleterious peptide expression: (a) plasmid copy number control by inactivation of the *pcnB* gene and (b) expression control through inactivation of the LacI inducer binding domain. In the third part of the thesis, I show that the positive random peptides confer competitive fitness advantage only under stressful conditions, for example, an elevated temperature. In conclusion, I show that random sequences indeed affect fitness of the host possibly through targeting specific genes or proteins. This study provides experimental evidence on how random sequences could serve as drivers of *de novo* gene evolution.

Table Of Contents

Chapter 1. Introduction

1.1. Mechanisms for producing novelty in genomes

Origin of novelty by adaptive evolutionary innovations is a fundamental theme in evolutionary genetics (reviewed in (Ding et al., 2012)). Genetic changes in existing genes can lead to innovation by affecting biological function and the roles these genes may play in the downstream pathways. It is now possible to compare entire genome sequences of different species and study the genomic features and changes that they have accumulated in the course of evolution. For example, variation in the number of genes between different closely related species allows for identification of new genes. Different mechanisms have been shown to be responsible for producing new genes, all of which contribute to evolutionary innovation.

1.1.1. Novelty using pre-existing genes as templates

It was hypothesized in the 1930s by Haldane and Muller (Haldane, 1932; Muller, 1936) that new genes may emerge from copying of pre-existing genes by the process of gene duplication. Gene duplication is by far the most widely studied mechanism, as it has been around since 1970s when Susumu Ohno described it in his book as the principal source of new gene evolution (Ohno, 2013). But what is the fate of these newly duplicated segments or genes? The basic model that was proposed is called the neo-functionalization model, wherein the copied gene that is now free of the constraints of natural selection (due to its redundant function), could accumulate mutations for a novel function (Kimura and Ohta, 1974). The newly duplicated gene can also serve to be redundant (i.e. to increase dosage of a gene product) or may be stabilized by sub-functionalization (divide the gene function into subfunction by each copy (Force et al., 1999)). A special case of new gene evolution was described in yeast species, whereby there is whole genome duplication or polyploidy, which is known to play role in producing novelty by dosage effects rather than new functions (Wapinski et al., 2007). Apart from DNA-mediated duplications, another well-known mechanism is retroposition or retroduplication, which can also give rise to new genes (reviewed in (Kaessmann et al., 2009)). Here, an mRNA of an existing gene is reverse transcribed into DNA and inserted back into the genome. Gene fusion is another mechanism that produces new genes, which occurs when two different genes fuse into one transcriptional unit and provide a novel function (reviewed in (Long et al., 2003)). Other mechanisms like exon shuffling have also shown to contribute to new functions (Kaessmann et al., 2002).

In bacteria such as *Escherichia coli*, different species can have large differences in their gene repertoire. Innovation may have played a major role in the generation of novelty in bacteria. The most common process known for this is lateral gene transfer (LGT), where DNA is transferred to the host from an external source, is known to substantially contribute to the genomic content of bacteria (Lawrence and Hendrickson, 2003; Lerat et al., 2005). In changing environments, LGT provides a great advantage to the bacteria, where they can acquire genes involved in transport and metabolism to cope with the changing nutrient supply. Template-based mechanisms (*i.e.* new genes evolving from pre-existing genes) of new gene evolution have been studied widely, although now there is also an accumulating body of evidence on the non-template mediated new gene evolution, which I describe in the upcoming section 1.1.2.

1.1.2. Non-template mediated novelty: *De novo* gene evolution

Non-template mediated novelty is when a pre-existing gene template is not required (e.g. from non-coding regions) for the evolution of new functional genes. That non-coding sequences can have biological functions was considered almost impossible, so much so, that François Jacob quoted in his famous essay, Evolution and Tinkering: “The probability that a functional protein would appear *de novo* by random association of amino acids is practically zero . . . creation of entirely new nucleotide sequences could not be of any importance in the production of new information” (Jacob, 1977). Indeed Jacob recognized that genes and proteins originated from somewhere, but he considered this to have happened previously in the prebiotic phase: “The really creative part in biochemistry must have occurred very early" (Jacob, 1977). But with the start of systematic sequencing efforts, it became clear that there are genes that cannot be traced back to a known ancestor (no homologues) and these were called orphan genes. This term was introduced in the late 1990s, where sequencing and annotation of the yeast genome showed that many protein coding open reading frames (ORFs) had no homologs in any annotated organism (reviewed in (Dujon, 1996)). As more and more genomes were sequenced a more systematic approach was developed to identify the emergence of genes, which was called phylostratigraphy (Domazet-Loso et al., 2007). Eventually it was realized that orphan genes are candidates for a true class of genes that could have evolved *de novo* out of non-coding sequences (Figure 1.1).

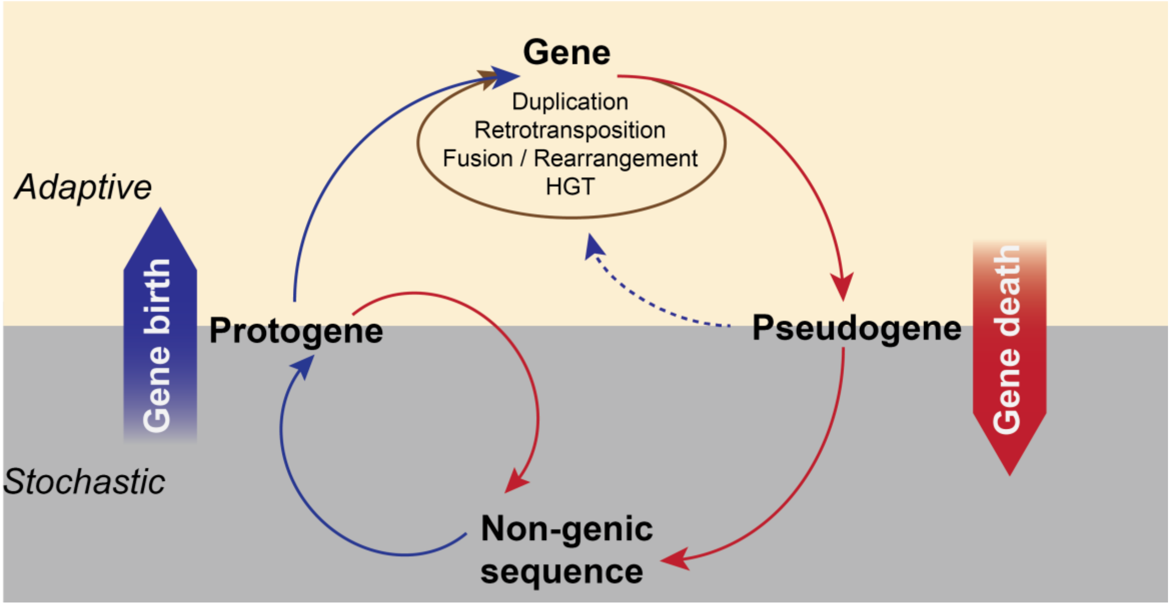

Figure 1.1. Emergence of new genes.

The cycle of genes shows the emergence of *de novo* genes (blue arrows) from a non-genic sequence with an intermediate of a 'protogene' (sequences that have gene-like properties; (Carvunis et al., 2012)). Red arrows show loss of function or pseudogenization which occurs often when there is lack of selective pressure in all of the mechanisms of gene evolution. Dotted blue arrow represents genes emerging from pseudogenes in the presence of selective positive pressure. Brown arrow shows template based processes by which new genes emerge (i.e. by using pre-existing genes). Image modified from (Neme and Tautz, 2014).

The first description of candidates for *de novo* genes was in *Drosophila* sp. (Begun et al., 2006; Levine et al., 2006). Subsequently, more candidates were found and it was shown that *de novo* genes can evolve rapidly from non-coding sequences

(Reinhardt et al., 2013). Several other comparative studies in yeast (Cai et al., 2008; Li et al., 2010), hydra (Khalturin et al., 2008), mouse (Heinen et al., 2009; Murphy and McLysaght, 2012), primates (Knowles and McLysaght, 2009; Wu et al., 2011; Xie et al., 2012) and plants (Xiao et al., 2009) have provided evidence of emergence of *de novo* genes. Today, there is enough evidence that new genes can originate from a previously non-genic sequence (reviewed in (Tautz and Domazet-Loso, 2011; Schlotterer, 2015)). Non-coding sequences also have been shown to be highly transcribed with the transcripts coding for small ORFs (Wade and Grainger, 2014) and these ORFs were found to be translated as shown by their association to ribosomes (Wilson and Masel, 2011).

Intergenic space from which *de novo* genes could evolve is rather limited in smaller organisms such as bacteria and viruses. Yet, these organisms have ways of innovating from non-coding templates. Bacteria have been shown to accumulate a large number of antisense transcripts, which may have important functions that are yet not known (Dornenburg et al., 2010). Another mechanism that exists in bacteria in order to generate novelty is known as overprinting (described in (Grassé, 2013)). Overlapping alternate open reading frames present in genes can become functional by creation of a promoter. Several genes have been described to have overprinting in (Delaye et al., 2008; Delaye et al., 2008; Fellner et al., 2014). Viruses also have the mechanism of overprinting (Keese and Gibbs, 1992; Sabath et al., 2012). Making use of overlapping ORFs allows for *de novo* evolution without having designated non-coding regions in prokaryotes.

1.2. Random sequences as a source of novelty

Non-coding sequences are largely random stretches of DNA between two genes or exons of the same gene. In the previous sections, I introduced non-coding sequences as the starting material for *de novo* gene evolution and provided several experimental studies that have shown this in the past. According to the Oparin-Haldane view on the origin of life, molecular complexity arose from long series of spontaneous steps, forming reproducing proto-cells (i.e. the first reproducing cells).

Analogously, random sequences can be viewed as the starting material, in the context of prebiotic evolution of the first biopolymers or specifically biopeptides. In order to be functional, a random sequence should translate into a polypeptide, ideally with a secondary structure capable of performing a function like catalysis or binding. The limited number of protein families that exist today are a result of a long selection process, which are believed to be constrained due to historical contingency (Chothia, 1992; Luigi Luisi, 2003). Previous studies using the phage display technique using a random peptide library have shown that it is possible to obtain as high as 20% random peptides with a functional fold (Chiarabelli et al., 2006). Other crucial *in vitro* studies from the early 1990s have screened for RNA and DNA molecules with specific functional properties (Bartel and Szostak, 1993; Famulok and Szostak, 1993). This raises questions like: What are the inherent properties of random DNA or peptide sequences? What functions could they adopt in organisms?

1.2.1. Properties of random protein sequences

In order to perform a function in an organism, random sequences should possess properties that allow for specific interactions with other molecules of the cell. It has been shown that random peptides have several inherent properties that place them only a few mutational steps away from the naturally existing protein families (Ptitsyn and Volkenstein, 1986). For example, it has been shown that random peptides can form 3D structures similar to those of globular proteins (Ptitsyn, 1985). On the other hand, random sequences could potentially show aggregation behaviour, which can lead to a substantial constraint for gain of function. In general, protein aggregation is deleterious to the cells, for example, protein misfolding leads to aggregation of disease causing amyloid fibrils (Dobson, 2003). If the random peptides have a higher aggregation propensity they might not provide any selective advantage due to the consequent decrease in fitness. This question was addressed in a previous computational study, in which the authors showed that aggregation propensity of random polypeptides is low (Angyan et al., 2012). With this they inferred that emergence of *de novo* peptides is not constrained by aggregation. As aggregation, another characteristic that has been widely studied in random sequences, is the intrinsic structural disorder (ISD). ISD determines the degree to which a given peptide folds into a stable 3D structure (ordered) versus an unstructured (disordered) structure. A higher ISD relieves evolutionary constraints, which allows for innovation (Romero et al., 1998; Mosca et al., 2012). Studies have shown that *de novo* proteins have a higher ISD, specifically young genes show a higher disorder compared to older genes (Wilson et al., 2017). Random sequences have an intrinsic ability to form structures, which theoretically are easily incorporated into the existing pool of genes and proteins. Empirical data have largely corroborated these studies, which I will introduce in the following section.

1.2.2. Functionality of random sequences in organisms

Functionality of random sequences has been shown in several studies that have made use of screening methods, where a specific selection pressure was employed. For example, early *in vitro* selection experiments have identified ribozymes (RNA enzymes) that catalyse various chemical transformations (Robertson and Joyce, 1990; Bartel and Szostak, 1993). Similar selection experiments were conducted using random sequence libraries where mRNA display technique was used to isolate novel ATP-binding proteins (Keefe and Szostak, 2001). These studies, although extremely relevant, were performed with a synthetic biology view. Albeit recently, there is accumulating evidence from studies in different organisms, which have systematically shown screening of functional random peptides under selective conditions.

Random sequences have shown to be functional as regulatory elements. Evolution of *de novo* promoters from random sequences has been shown previously by replacing the -35 promoter region of the tetracycline resistance gene on a plasmid with random bases followed by screening of resistant clones (Horwitz and Loeb, 1986). Another study used random sequence libraries to replace the entire lac promoter region (103 bp) in (Yona et al., 2018). It was shown that 60% of random promoters restored the lac operon expression in just one mutational step and about 10% variants resorted function after two mutational steps. It is not surprising that the authors found most mutations targeting -10 and -35 regions of the Pribnow box (required for initiation of transcription). Other studies in bacteria such as *E. coli* have used directed evolution approach to select for novel random peptides that conferred resistance to certain antibiotics (Knopp et al., 2019) or certain biotoxic agents (Stepanov and Fox, 2007). With similar approaches, several *de novo* proteins have been discovered in *E. coli* that confer resistance to metal toxicity (Hoegler and Hecht, 2016; Hoegler and Hecht, 2018). Small random peptides expressed in *Arabidopsis thaliana* plants have shown that specific novel peptides affect phenotypes like photosynthesis, flowering time and red-light response (Bao et al., 2017). The authors showed that certain random peptides have the ability to interfere or enhance biological processes. Most studies (except Bao et al.) until now have focused on a screening by a selection approach, where restrictive growth conditions have been used to select for phenotypic rescue. In our lab, a previous study has used optimal growth conditions, to show bioactivity of a random peptide coding library in (Neme et al., 2017). My thesis is a continuation of this study and hence, I will describe it in more detail in the subsequent section.

1.3. Bioactivity of random sequences: A proof of concept study

Emergence of new genes by *de novo* evolution can be mimicked by using random stretches of nucleotides to test for functionality. The process can be accelerated if experiments are done at the 'protogene' stage (Figure 1.1), where one provides all necessary elements for successful transcription and translation of random sequences and then look for functions. This idea was tested in our group, where authors provided randomly synthesized coding library (akin to protogenes) on an inducible vector to study their effects in (Neme et al., 2017).

A library consisting of oligonucleotides made from random sequences (synthesized in equimolar concentrations of A, T, G, and C) engineered into an inducible plasmid vector (see 2.1.1) was designed (Figure 1.2A). The inserts were designed such that they had the necessary elements for transcription (TATA box) and translation (start codon, ribosome binding site (RBS) and stop codon with a FLAG tag for peptide detection) flanking them (Figure 1.2B). This was done to allow the expression of random sequences inside the cells upon induction, the artificial protogene. The frequency of each variant clone could easily be tracked over time using deep sequencing of the plasmid DNA, wherein each unique random sequence acted like barcodes. In two parallel sets of experiments, an overnight culture of containing the transformed library (10 replicates), was serially propagated through four cycles of induction with 1 mM IPTG for a time span of 3 or 24 hours respectively. Control was the non-induced library. Large numbers of bacteria were transferred from one cycle to the next to avoid bottlenecks where frequencies could have changed purely due through effects of drift. After each cycle (i.e. 3 hours or 24 hours) just before passaging, plasmid DNA was isolated and sequenced to determine the frequency of each random sequence from the transformed library database (screening-by-sequencing approach). Since clonal populations of were used, any changes were assumed to be due to a competitive growth outcome.

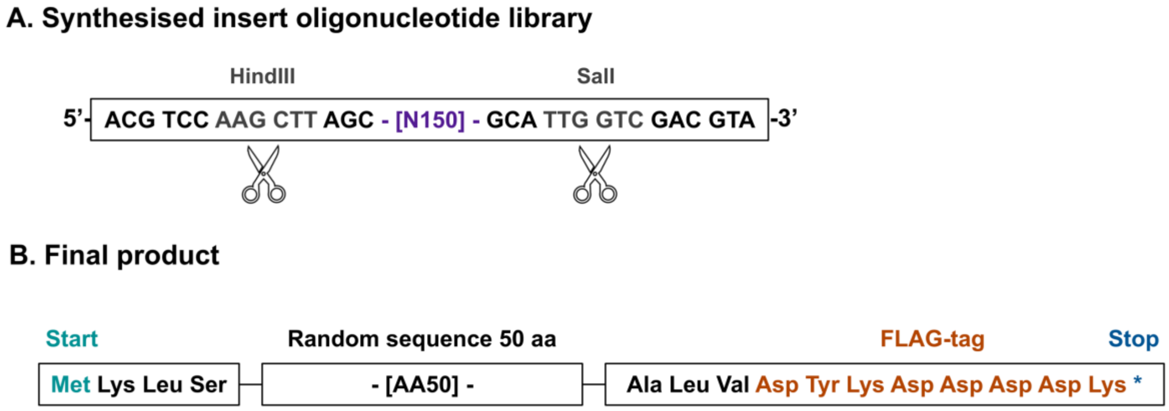
A. Synthesised insert oligonucleotide library
HindIII
SalI
5'- ACG TCC AAG CTT AGC - [N150] - GCA TTG GTC GAC GTA -3'
B. Final product
Start
Random sequence 50 aa
FLAG-tag
Stop
Met Lys Leu Ser
- [AA50] -
Ala Leu Val Asp Tyr Lys Asp Asp Asp Asp Lys *

Figure 1.2. Random coding library design.

A) Insert of 150 random A, T, G and C nucleotides synthesized with common flanking sites consisting of restriction sites HindIII and SalI for ligation into the vector pFLAG-CTC. B) Final product that is produced after IPTG induction. The in-frame start site, FLAG-tag and the stop site are from the vector backbone.

The experiments revealed abundant changes in the frequency of random sequence variants over time (Figure 1.3). A fraction of variants went down in frequency, whereas a fraction of sequences increased in frequency after four cycles under

IPTG induction (red and green circles, Figure 1.3A). On the other hand, no significant changes were seen in the non-induced populations, where the random peptides were not expressed (in black, Figure 1.3B). The changes in frequency meant that clones carrying a particular random sequence grew either better or worse under competition, compared to the millions of variant clones present in the population. These sequences were later classified as: deleterious (decreased in frequency), neutral (no significant change in frequency) and non-deleterious/beneficial (increased in frequency). In my thesis, I test for selected candidates from all of the aforementioned categories and assess the properties and biological significance of those in model bacterium Escherichia coli.

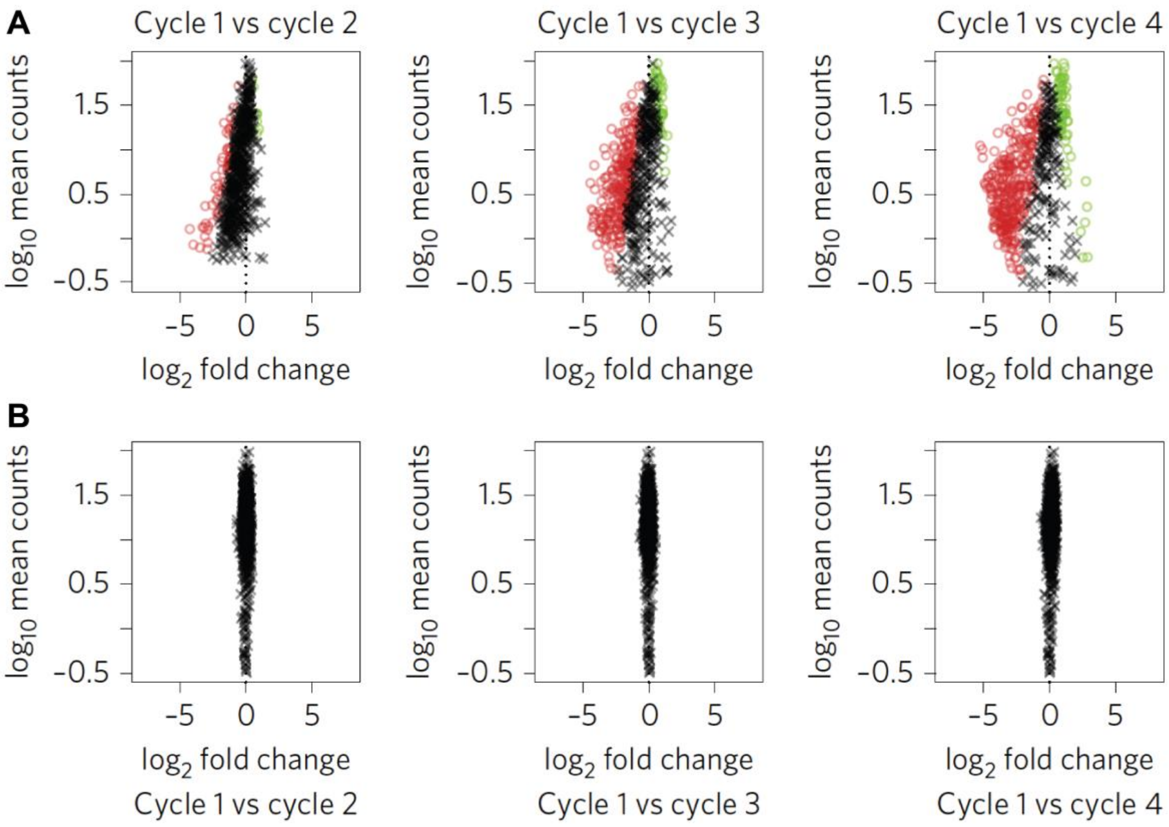

Figure 1.3. Expression of random sequence library in drives changes in individual variant frequency over time.

A) Induced B) Non-induced control. Plots show fold change versus mean counts of each sequence in the library passaged for four days (24 hours/cycle). The first cycle was compared with the 2nd, 3rd and 4th cycle to track frequency changes. Green and red dots indicate positive and negative fold changes respectively. Image modified from (Neme et al., 2017).

1.4. Aims of this study

Random peptides have been shown to have biological functions in different organisms. Directed evolution experiments have shown that functional random peptides can be screened under selective conditions. Previous study from our lab has shown that random peptides show bioactivity under optimal growth conditions after few serial passages in *E. coli* (Neme et al., 2017). Change in the frequencies of different random sequence variants provided evidence that every sequence variant affected the host uniquely, causing changes in growth rates of the host carrying particular variants.

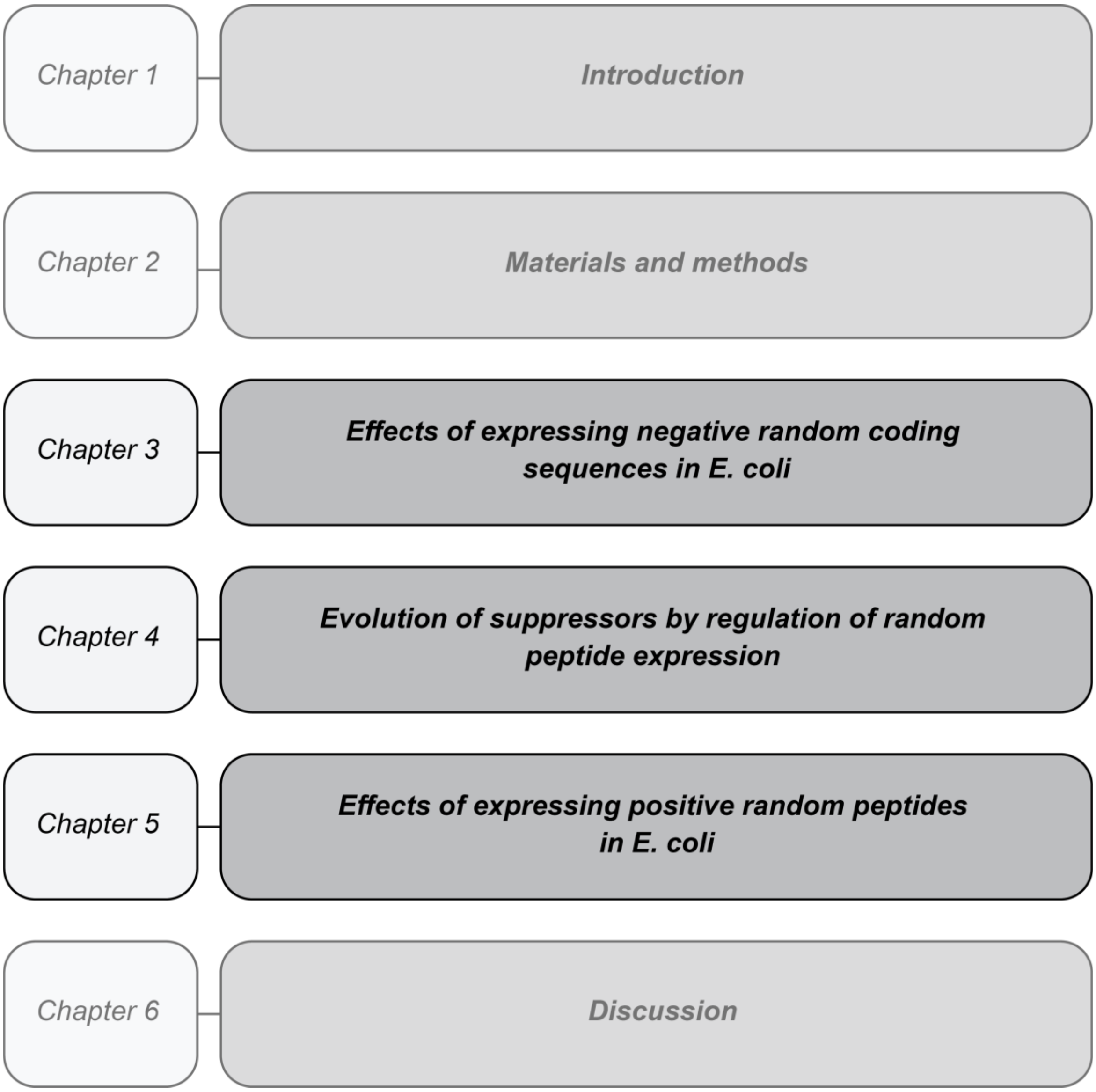
Chapter 1
Introduction
Chapter 2
Materials and methods
Chapter 3
Effects of expressing negative random coding sequences in E. coli
Chapter 4
Evolution of suppressors by regulation of random peptide expression
Chapter 5
Effects of expressing positive random peptides in E. coli
Chapter 6
Discussion

Figure 1.4. Aims of this thesis.

Three chapters of this thesis are dedicated to the findings. Chapter 1 is introduction, chapter 2 is materials and methods and finally, chapter 6 consists of the general discussion.

The main theme of this thesis is to understand the interaction and effects of different clones individually in the model bacterium, *Escherichia coli*. The effects of different random sequence clones are described in detail with an aim to understand individual effects of random peptides in *E. coli*. The findings are divided into three chapters as shown in Figure 1.4. I describe the effects of deleterious random peptides with an attempt to uncover possible cellular interactions of the peptides with the host machinery (Chapter 3). The mutational targets specifically affecting expression of candidate peptides are described in Chapter 4, to recognize the various coping mechanisms used by host bacteria. In Chapter 5, I aim to understand effects of individual positive peptides on the fitness of *E. coli*.

Chapter 2. Materials and methods

2.1. Materials

2.1.1. Bacterial strains and plasmids

Bacteria were grown overnight (16-18 h) followed by preparation of glycerol stocks. Stocks were made by taking 1:1 volume of the stationary phase culture and 50% w/v of glycerol. The mixture was flash frozen in liquid nitrogen and stored at -70°C indefinitely. To avoid freeze thaw cycles, cultures were scrapped off from the frozen vials without allowing them to thaw (by keeping tubes in dry ice) and used for streaking on the desired medium for revival prior to experiments.

All strains are *Escherichia coli*

Strain name	Markers	Obtained from
K-12 DH10B derivative (Neb 10-beta)	F$^-$, endA1, deoR$^+$, recA1, galE15, galK16, nupG, rpsL, Δ(lac)X74, φ80lacZΔM15, araD139, Δ (ara-leu)7697, mcrA, Δ(mrr-hsdRMS-mcrBC), Str^R, λ$^-$	New England Biolabs, Cat. No. C3019H
K-12 MG1655	K-12, F$^-$, λ$^-$, ilvG$^-$, rfb-50, rph-1	Dr. Jenna Gallie*
B REL606	F$^-$, tsx-467(Am), lon$^-$, araA230, rpsL227(strR), hsdR$^-$, [mal$^+$](LamS)	Dr. Jenna Gallie*
B REL607	F$^-$, tsx-467(Am), lon$^-$, rpsL227(strR), hsdR$^-$, [mal$^+$](LamS)	Dr. Jenna Gallie*
K-12 BW25113 (Parent strain of Keio collection)	Δ(araD-araB)567, λ$^-$, ΔlacZ4787(::rrnB-3), rph-1, Δ(rhaD-rhaB)568, hsdR514	The coli genetic stock centre (Keio strains)
K-12 JW5808-1	Δ(araD-araB)567, λ$^-$, ΔpcnB759::kan, rph-1, ΔlacZ4787(::rrnB-3), Δ(rhaD-rhaB)568, hsdR514	The coli genetic stock centre (Keio strains)

Table 2.1. Description of the bacterial strains used.

All strains used in this study were *Escherichia coli* sp. obtained from various sources as mentioned. *The sources mentioned are not necessarily where the strains originated but rather from where I personally obtained them (in the form of a subculture, here, kindly from Dr. Gallie).

Plasmids were stored as glycerol stocks after transforming into the desired strain backgrounds mentioned in Table 2.1. Plasmids used in this study are mentioned in Table 2.2. Plasmid maps with features is provided in the Figure 2.1.

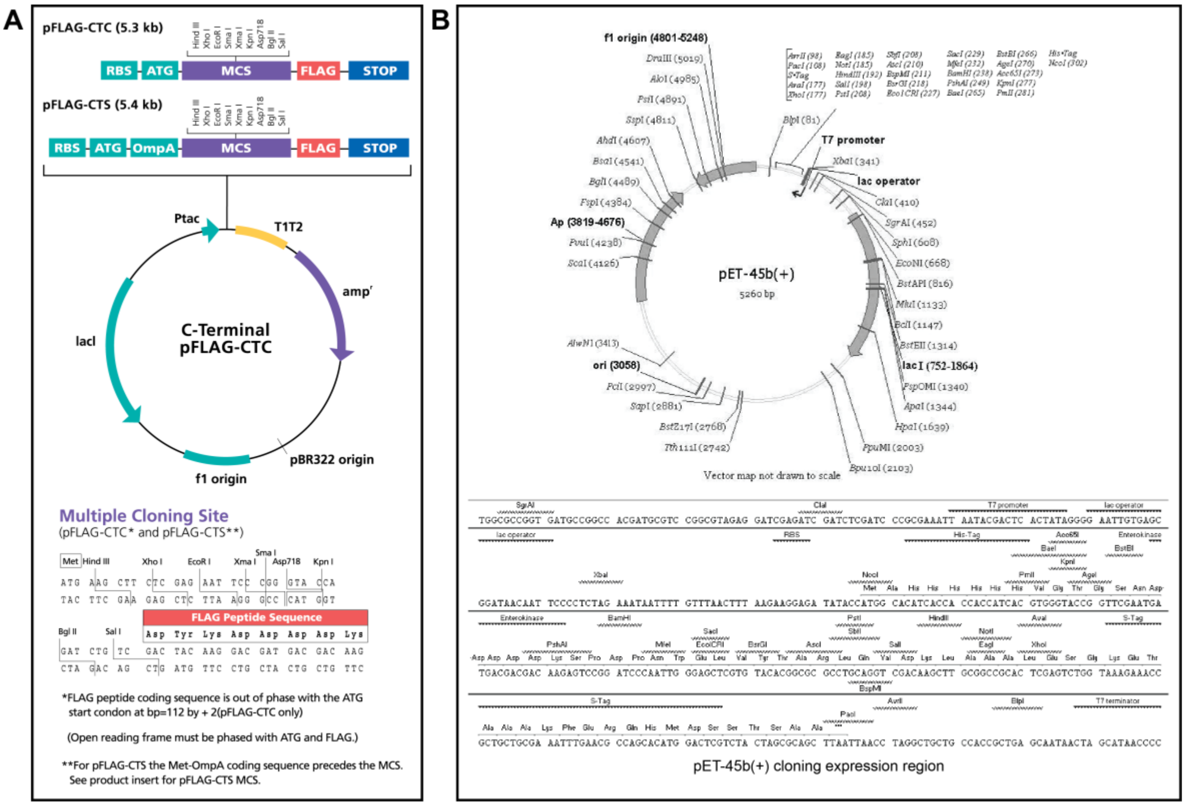
A
pFLAG-CTC (5.3 kb)
Hind III
Xho I
EcoR I
Sma I
Xma I
Kpn I
Asp718
Bgl II
Sal I
RBS
ATG
MCS
FLAG
STOP
pFLAG-CTS (5.4 kb)
OmpA
Ptac
T1T2
amp'
lacI
C-Terminal pFLAG-CTC
pBR322 origin
f1 origin
Multiple Cloning Site
(pFLAG-CTC* and pFLAG-CTS**)
Met
ATG AAG CTT CTC GAG AAT TCC CGG GTA CCA
TAC TTC GAA GAG CTC TTA AGG GCC CAT GGT
FLAG Peptide Sequence
Asp Tyr Lys Asp Asp Asp Asp Lys
GAT CTG TC GAC TAC AAG GAC GAT GAC GAC AAG
CTA GAC AG CTG ATG TTC CTG CTA CTG CTG TTC
*FLAG peptide coding sequence is out of phase with the ATG start condon at bp=112 by + 2(pFLAG-CTC only)
(Open reading frame must be phased with ATG and FLAG.)
**For pFLAG-CTS the Met-OmpA coding sequence precedes the MCS. See product insert for pFLAG-CTS MCS.
B
f1 origin (4801-5248)
DraIII (5019)
AloI (4985)
PsiI (4891)
SspI (4811)
AhdI (4607)
BsaI (4541)
BglI (4489)
FspI (4384)
Ap (3819-4676)
PvuI (4238)
ScaI (4126)
AlwN1 (3413)
ori (3058)
PciI (2997)
SapI (2881)
BstZ17I (2768)
Tth111I (2742)
AvrII (98) PacI (108) S•Tag AvaI (177) XhoI (177)
BagI (185) NotI (185) HindIII (192) SalI (198) PstI (208)
SfiI (208) AscI (210) BspMI (211) BsrGI (218) Eco1CRI (227)
SacI (229) MfeI (232) BamHI (238) PshAI (249) BaeI (265)
BstBI (266) AgeI (270) Acc65I (273) KpnI (277) PmlI (281)
His•Tag NcoI (302)
BlpI (81)
T7 promoter
XbaI (341)
lac operator
ClaI (410)
SgrAI (452)
SphI (608)
EcoNI (668)
BstAPI (816)
MluI (1133)
BclI (1147)
BstEII (1314)
lacI (752-1864)
PspOMI (1340)
ApaI (1344)
HpaI (1639)
FpuMI (2003)
Bpu10I (2103)
pET-45b(+)
5260 bp
Vector map not drawn to scale
SgrAI ClaI T7 promoter lac operator
TGGCGCCGGT GATGCCGGCC ACGATGCGTC CGGCGTAGAG GATCGAGATC GATCTCGATC CCGCGAAATT AATACGACTC ACTATAGGGG AATTGTGAGC
lac operator RBS His-Tag Acc65I Enterokinase BaeI BstBI KpnI XbaI NcoI PmlI AgeI
Met Ala His His His His His His Val Gly Thr Gly Ser Asn Asp
GGATAACAAT TCCCCTCTAG AAATAATTTT GTTTAACTTT AAGAAGGAGA TATACCATGG CACATCACCA CCACCATCAC GTGGGTACCG GTTCGAATGA
Enterokinase BamHI PstI HindIII AvaI S-Tag SacI SbfI NotI PshAI MfeI Eco ICRI BsrGI AscI SalI EagI XhoI
Asp Asp Asp Asp Lys Ser Pro Asp Pro Asn Trp Glu Leu Val Tyr Thr Ala Arg Leu Gln Val Asp Lys Leu Ala Ala Ala Leu Glu Ser Gly Lys Glu Thr
TGACGACGAC AAGAGTCCGG ATCCCAATTG GGAGCTCGTG TACACGGCGC GCCTGCAGGT CGACAAGCTT GCGGCCGCAC TCGAGTCTGG TAAAGAAACC
BspMI
S-Tag AvrII BlpI T7 terminator PacI
Ala Ala Ala Lys Phe Glu Arg Gln His Met Asp Ser Ser Thr Ser Ala Ala
GCTGCTGCGA AATTTGAACG CCAGCACATG GACTCGTCTA CTAGCGCAGC TTAATTAACC TAGGCTGCTG CCACCGCTGA GCAATAACTA GCATAACCCC
pET-45b(+) cloning expression region

Figure 2.1. Vector maps as provided by respective the companies.

A) pFLAG-CTC is an expression vector with a strong promoter and a high copy number (>100 copies/cell; purchased from Sigma-Aldrich®). B) The pET45b (+) is also an expression vector (purchased from Novagen®) but has a copy number of about 20/cell (medium).

Plasmid names	Markers	Obtained from
pFLAG-CTC™ Size = 5.3 kb	Amp^R, pBR322 *ori,* f1 *ori*, *lacI,* Ptac promoter, C-Terminal FLAG-tag	Sigma-Aldrich®

pET45b (+) Size = 5.26 kb	Amp^R, pBR322 *ori,* f1 *ori, lacI,* T7 promoter, N- Terminal His-Tag	Novagen®

Table 2.2. Description of plasmids used in this study.

Plasmids were transformed into desired backgrounds as controls against plasmids with specific inserts which are listed elsewhere.

2.1.2. Candidate random sequences used in the study

This study includes a list of inserts which were cloned into the plasmids mentioned in Table 2.2 in combination with the strain backgrounds (Table 2.1). Following is a list of sequences used in this study and their genome and plasmid backgrounds. All of the genetically modified strains were stored in 50% glycerol at -70°C using the same method described in the section 2.1.1.

Insert name (Alias used in this study)	Nucleotide sequence of the coding inserts
Pep159 (NEG_Pep1)	ATGAAGCTTAGCGTTGGGAAACCGGATACATGGCTCCATAGAGCCAGAGGAGCAGTTTGGGTTAGGATTGCCGGGAACGTGTCATTGGGTATGGGACTGCGTGGTTTGTCTGGTCGGCTCCCATGTGTATGTGGGCCTCTCAGGACCTTTGGGGCCTTCGAGGCATTGGTCGACTACAAGGACGATGACGACAAG
Pep159-Stop (NEG_Pep1_Stop)	ATGAAGCTTTAGGTTGGGTAACCGGATACATGGCTCCATAGAGCCAGAGGAGCAGTTTGGGTTAGGATTGCCGGGAACGTGTCATTGGGTATGGGACTGCGTGGTTTGTCTGGTCGGCTCCCATGTGTATGTGGGCCTCTCAGGACCTTTGGGGCCTTCGAGGCATTGGTCGACTACAAGGACGATGACGACAAG
Pep292 (NEG_Pep2)	ATGAAGCTTAGCGCGGCTACCTGGGTCGCGAGTCTCCGAGTTGCCTTCGGTGGGGACCTTATTCTGCGGTTAATCAGATATCAGGCGGCAGGGCGAAGCGGAGCGCTCGACCAGTTTTATGAAGCGAACTCCATACTAGGTGTCCACAGGCGTACGCGAGATGCATTGGTCGACTACAAGGACGATGACGACAAG
Pep292-Stop (NEG_Pep2_Stop)	ATGAAGCTTTAGGCGGCTTAATGGGTCGCGAGTCTCCGAGTTGCCTTCGGTGGGGACCTTATTCTGCGGTTAATCAGATATCAGGCGGCAGGGCGAAGCGGAGCGCTCGACCAGTTTTATGAAGCGAACTCCATACTAGGTGTCCACAGGCGTACGCGAGATGCATTGGTCGACTACAAGGACGATGACGACAAG
Pep419 (NEG_Pep3)	ATGAAGCTTAGCTGTCCATTTCCGGATACCCATGGCGCGATCTGCTGCCGTGTTTGGTCCGGGTTCGCGTTGATTGTTCTGCGTTTGCTCGACGCCATAGGGTCCTGCGGCAGGCATGGTGGTGTG

Insert name (Alias used in this study)	Nucleotide sequence of the coding inserts
	GGCCACGCTCTAGCCGAGCACGTCTTCTGGGTGTGTGCATTGGTCGACTACAAGGACGATGACGACAAG
Pep419-Stop (NEG_Pep3_Stop)	ATGAAGCTTTAGTGTCCATAACCGGATACCCATGGCGCGATCTGCTGCCGTGTTTGGTCCGGGTTCGCGTTGATTGTTCTGCGTTTGCTCGACGCCATAGGGTCCTGCGGCAGGCATGGTGGTGTGGGCCACGCTCTAGCCGAGCACGTCTTCTGGGTGTGTGCATTGGTCGACTACAAGGACGATGACGACAAG
Pep555 (NEG_Pep4)	ATGAAGCTTAGCGTGTATATTCTTACGGTCCAGTTCTGCACTGGCTGGGGGGTGCCGATGGCCACGACATACTTGTATGCTGGGGGGCTGCGGCGGGGTCATCACCAAGGCCGCTCTGAGTCTTCTTATCGGAGTTTTCGTAAGCGTCGGGCTAACACGCTGGCATTGGTCGACTACAAGGACGATGACGACAAG
Pep555-Stop (NEG_Pep4_Stop)	ATGAAGCTTTAGGTGTATTAACTTACGGTCCAGTTCTGCACTGGCTGGGGGGTGCCGATGGCCACGACATACTTGTATGCTGGGGGGCTGCGGCGGGGTCATCACCAAGGCCGCTCTGAGTCTTCTTATCGGAGTTTTCGTAAGCGTCGGGCTAACACGCTGGCATTGGTCGACTACAAGGACGATGACGACAAG
Pep628 (NEG_Pep5)	ATGAAGCTTAGCTCAGTTTGCATCCTTGTCCTGGTTCTGAGGCACCGGTTAGACGCGTTGTGGCTAAGATTACGCTCGGAAGGGGCGATTAGCATCTTCAGTTGGCATGAGGAGAGCTATCGGGTCGGTGGCGATCTGTGCACAGAGCGCAAGCCCTCTCGGGCATTGGTCGACTACAAGGACGATGACGACAAG
Pep628-Stop (NEG_Pep5_Stop)	ATGAAGCTTTAGTCAGTTTAAATCCTTGTCCTGGTTCTGAGGCACCGGTTAGACGCGTTGTGGCTAAGATTACGCTCGGAAGGGGCGATTAGCATCTTCAGTTGGCATGAGGAGAGCTATCGGGTCGGTGGCGATCTGTGCACAGAGCGCAAGCCCTCTCGGGCATTGGTCGACTACAAGGACGATGACGACAAG
Pep629 (NEG_Pep6)	ATGAAGCTTAGCAAAGTAGTTTATCGTCGCGCAGCTCAGTCCCGTGCTCGGTCCGGCGGCTTGACCGGGGGTCGCGTGGAGGAAAATGATGTCCTTACGGGTGCGAGGGTGAGATTACGGGCTTTACTTTGTTGCGCGGGAGTCAGTGTCTGTGTAACCTCGGCATTGGTCGACTACAAGGACGATGACGACAAG
Pep629-Stop (NEG_Pep6_Stop)	ATGAAGCTTTAGAAAGTATAATATCGTCGCGCAGCTCAGTCCCGTGCTCGGTCCGGCGGCTTGACCGGGGGTCGCGTGGAGGAAAATGATGTCCTTACGGGTGCGAGGGTGAGATTACGGGCTTTACTTTGTTGCGCGGGAGTCAGTGTCTGTGTAACCTCGGCATTGGTCGACTACAAGGACGATGACGACAAG
Fam600T (POS_Pep1)	ATGAAGCTTAGCCGCGGTATTCACCTAGGTCGGACGAGTACATGCGTCAACGCTTCGTACGCACTCTGCCACACGTACCGTTCAGCCCGCCGTGGCAAGTCCAGGAAGAGTGGGAGGAGTTCACCACCGATCGGGACCTCTTTAGTACACTGGGTTTTGGACGCATTGGTCGACTACAAGGACGATGACGACAAG
Fam600T-Stop (POS_Pep1_Stop)	ATGAAGCTTAGCCGCGGTATTCACCTAGGTCGGACGAGTACATGCGTCAACGCTTCGTAGGCACTCTGCCACACGTACCGTTCAGCCCGCCGTGGCAAGTCCAGGAAGAGTGGGAGGAGTTAACCACCGATCGGGACCTCTTTAGTACACTGGGTTTTGGACGCATTGGTCGACTACAAGGACGATGACGACAAG

Insert name (Alias used in this study)	Nucleotide sequence of the coding inserts
Fam32 (POS_Pep2)	ATGAAGCTTAGCTACTGGAATAGCTCTATGGCGTCGGGGGATATCCGTGCTCTTGTGTTTGATTCAGGCGGAGGCTTAATATTCCTTCGGCATCAGCTGGCGGGGTGGTGGGCCTGTTTGTTTCCGCTACTGGCATCGCGGGAGGCACGGTTTGATACCGACGCATTGGTCGACTACAAGGACGATGACGACAAG
Fam32-Stop (POS_Pep2_Stop)	ATGAAGCTTAGCTGATGGAATAGCTCTATGGCGTCGGGGGATATCCGTGCTCTTGTGTTTGATTCAGGCGGAGGCTTAATATTCCTTCGGCATCAGCTGGCGGGGTGGTGGGCCTGTTTGTTTCCGCTACTGGCATCGCGGGAGGCACGGTTTGATACCGACGCATTGGTCGACTACAAGGACGATGACGACAAG
Fam4 (POS_Pep3)	ATGAAGCTTAGCCCCGTCTCCTGGATTCACGGTGCTACCGCTCAGTCTGGAGGATTATCCCTCAGGCTTGCAGTCCGCTCAGGAATAGATGGGTGTGCATGGTTCATCAGGGCTGAATGCGGAGGGGCTCGTGCGCTTTCAGACGGGCCTGGGGTAAGCTATGCATTGGTCGACTACAAGGACGATGACGACAAG
Fam4-Stop (POS_Pep3_Stop)	ATGAAGCTTAGCTGAGTCTCCTGGATTCACGGTGCTACCGCTCAGTCTGGAGGATTATCCCTCAGGCTTGCAGTCCGCTCAGGAATAGATGGGTGTGCATGGTTCATCAGGGCTGAATGCGGAGGGGCTCGTGCGCTTTCAGACGGGCCTGGGGTAAGCTATGCATTGGTCGACTACAAGGACGATGACGACAAG
HFC_PosA (POS_Pep4)	ATGAAGCTTAGCGTCATGCGTCCCATATCTCGCACCCTCCAGGTTGGTTATAACGGTCGCTGCGGCCAGTGCACTCAGCCGTTAAACCCACCCTCGTGCATCTTGGACACTTCGCCCGGGGGCTTGTGGACAGCAAGTTTTGCTCGTGGTAACAATCGCGAAGCATTGGTCGACTACAAGGACGATGACGACAAG
PosA-Stop (POS_Pep4_Stop)	ATGAAGCTTAGCGTCTAGCGTCCCATATCTCGCACCCTCCAGGTTGGTTATAACGGTCGCTGCGGCCAGTGCACTCAGCCGTTAAACCCACCCTCGTGCATCTTGGACACTTCGCCCGGGGGCTTGTGGACAGCAAGTTTTGCTCGTGGTAACAATCGCGAAGCATTGGTCGACTACAAGGACGATGACGACAAG
HFC_PosC (POS_Pep5)	ATGAAGCTTAGCGAAGGTGGCCGCCGATGTCTATGCACGGAAGCGACGGGGCTCTTTGCTGCCCTGTGCGGGCGTACGCCGGTTTATTTCAGTGTCGTATGCGGTCCATCCTGCATGTCGTTTGAATGCGGTCATTGTAGGCGTCTAACTTTATGCCGTACCGCATTGGTCGACTACAAGGACGATGACGACAAG
PosC-Stop (POS_Pep5_Stop)	ATGAAGCTTAGCTAGGGTGGCCGCCGATGTCTATGCACGGAAGCGACGGGGCTCTTTGCTGCCCTGTGCGGGCGTACGCCGGTTTATTTCAGTGTCGTATGCGGTCCATCCTGCATGTCGTTTGAATGCGGTCATTGTAGGCGTCTAACTTTATGCCGTACCGCATTGGTCGACTACAAGGACGATGACGACAAG
NT76 (NT_Pep1)	ATGAAGCTTAGCTCAGTCGATTGGCGGGCTTCAAGTAAATGGTCCATCCTTCTGCACTGGGGAGCCCAGGCGATGATGAACTATTCGACGGGCATGCGCGTCGCTTATGACCGGCGCGGCTTCAGGAGGACGCCCTTCGGCGCGGCTCATCCTGATCCTTTCGCATTGGTCGACTACAAGGACGATGACGACAAG
NT596	ATGAAGCTTAGCTCTTGGGCAGCGAAACCACAGAACGAGCTTCAAGCAAGGGGTTACGAAGATGTGTGCAAGGCCTACTTAATGTTGGCCCATGGGGGCAACGAACTGCGCTGCTGCACAGCGGCA

Insert name (Alias used in this study)	Nucleotide sequence of the coding inserts
(NT_Pep2)	CTGCGGGCGCTTTTGCGCGCGGCGCTTAGGATAGTCGCATTGGTCGACTACAAGGACGATGACGACAAG
NT1683 (NT_Pep3)	ATGAAGCTTAGCCAGCTCTCGTTCGTCATAGTCAATGGACCTTGTACTGCGACTAGCGTTTCGGCTCCCCGTTCGCTGTGTGAACCGAGTCGCAGCGTTGGCGGCTCGTTTATACGCTGGACTTGTTACGGCATCCATGGCGCGAAGTGCGACAATTCTCCTGCATTGGTCGACTACAAGGACGATGACGACAAG
NT1757 (NT_Pep4)	ATGAAGCTTAGCCCCAAGCCTCGGTGGCAGCTGGTATCGGATATGATGGTTACAGGGGTATCGCGCGAGAGTTTTGAGACGCCGGGCGGCATAAATCGTTGGCTACGTAGCGAGGCACTGCGCTTTGGGCATAAAACTGGAGGGATGTTATTAACGGGTTTGGCATTGGTCGACTACAAGGACGATGACGACAAG
NT2613 (NT_Pep5)	ATGAAGCTTAGCCCTTCTCTCGAGTTTGGGCTAAAAGCGGTCCATATGAGTGCCCACGCGCGCCGTGCAGAACCGGCGGGACGCGCGTATAAGTCGATAACGCAGGGGCACGTGGTGTATGAAAGGGGTCCTTTTCGAGCCGGACGCCATATCGTGCAGGACGCATTGGTCGACTACAAGGACGATGACGACAAG

Table 2.3. Nucleotide sequence list of selected candidates used for cloning.
List of all candidates used in this study are listed above. In-frame stop codons of the control sequences are underlined in maroon. Cloning in pFLAG-CTC and pET45b (+) followed by transformation in various backgrounds mentioned in Table **2.1** was performed to obtain modified strains expressing candidate peptides.

2.1.3. Growth media and culture conditions

Bacteria were cultured in Luria Bertani (LB) Lennox broth medium containing 10 g/L tryptone, 5 g/L yeast extract and 5 g/L NaCl or in minimal medium (M9-Glucose) containing 33.9 g/L NaH_2PO_4, 15 g/L of KH_2PO_4, 2.5 g/L NaCl, 5 g of 1.8 M NH_4Cl, 50 µL 1M $CaCl_2.6H_2O$, 1 mL 1 M $MgSO_4.7H_2O$ and 10 mL 20% glucose. Glucose was filter sterilized separately and then added to the rest of the autoclaved M9 media to prevent charring of sugar. Bacteriological agar (1.5% w/v) was added to the mentioned liquid media when solid media plates were required (for streaking stock cultures, plating etc.). Cultures from glycerol stocks were streaked on appropriate growth media plates and incubated at 37°C overnight (unless otherwise stated) in order to obtain single isolated colonies. These single colonies were then used as individual replicates to inoculate 4 mL of appropriate liquid medium in 14 mL plastic tubes with caps (Corning; Cat. No. 352057) and incubated at 37°C with a shaking speed of 250 rpm for 16-18 h (unless otherwise mentioned) to obtain an early stationary phase culture. Tetrazolium agar (TA) plates containing 10 g/L tryptone, 1 g/L yeast extract, 5 g/L NaCl, 16 g/L agar, 10 g/L L-arabinose and 1 mL 5% Tetrazolium indicator dye (TTC) were used for distinguishing Ara^+/Ara^- strains during competition experiments using B REL606/REL607 strain backgrounds. Arabinose sugar and TTC were separately autoclaved and filter sterilized respectively and added later to the rest of the autoclaved media. Ringer's solution containing 0.12 g/L $CaCl_2$, 0.105 g/L KCl, 0.05 g/L $NaHCO_3$ and 2.25 g/L NaCl was used for preparing serial dilutions of bacterial cultures. The solution maintains osmotic balance of the bacteria (Sigma; Cat. No. 96724). All chemicals were purchased from Sigma-Aldrich (now Merck) unless otherwise mentioned.

2.1.4. Antibiotics used in the study

Antibiotics were used to maintain the plasmids in all the *E.coli* background strains. Ampicillin was stored as a 50 mg/mL stock (dissolved in water) solution and was used as a final concentration of 50 μg/mL for both solid and liquid media. TA plates were made by adding 50 μg/mL of ampicillin and 100 μg/mL of streptomycin (stored as 100 mg/mL stock concentration) dissolved in water. For long term storage, all antibiotic stocks were stored at -20°C in 1 mL aliquots (to avoid repeated freeze thaw cycles).

2.1.4.1. Isopropyl β-d-1-thiogalactopyranoside (IPTG) for plasmid induction

All plasmids were inducible with the addition of IPTG. 1 M IPTG was made my dissolving 23.8 g in 100 mL distilled water followed by filter sterilization. Aliquots of 1 mL were stored at -20°C to avoid frequent freeze-thaw cycles. Working concentration that was used to induce plasmids was always 1 mM.

2.1.5. Antibodies used

Western blots were performed in order to check the expression of the peptides post induction with IPTG. After transferring the proteins on a PVDF membrane (section 2.2.10), the membranes were stained with a monoclonal anti-FLAG primary antibody, to target the FLAG tagged peptides of interest followed by a polyclonal secondary antibody (Table 2.4), used to recognize the primary antibodies bound to target.

Antibody	Produced in	Type	Catalogue No.
Anti-FLAG M2	Mouse	Monoclonal	GenScript: GENSA00187
IgG Alkaline phosphatase	Goat anti-mouse	Polyclonal	Sigma: A3562

Table 2.4. Antibodies used for western blotting.
The primary antibody was stored as 10 µL aliquots at -20°C to avoid freeze-thaw cycles and the secondary antibody was stored as is at 4°C.

2.1.6. Colorimetric detection of western blots

After staining the membranes with primary and secondary antibodies, a colorimetric detection was performed. The membrane was stained with 35 µL of nitro-blue tetrazolium chloride (NBT: Merck 11383213001) with 45 µL of 5-bromo-4-chloro-3'-indolphosphate p-toluidine salt (BCIP: Merck 11383221001) solutions mixed in 10 mL of 1X detection solution (DIG – from the DIG wash and block buffer set: Merck 11585762001) and incubated in the dark (Section 2.2.10).

2.1.7. Primers

Primers were synthesized by Sigma-Aldrich and provided as a lyophilized powder by the company. Upon arrival, primers were suspended in ultrapure distilled water (Invitrogen; Cat. No. 10977) to a concentration of 100 pmol/µL. All working primer stocks were stored at -20°C at a concentration of 10 pmol/µL.

Name	Sequence (5'→ 3')
Primers for the pFLAG-CTC vector inserts	
159_FWD	ATGAAGCTTAGCGTTGGGAAACCGGATACATGGC
292_FWD	ATGAAGCTTAGCGCGGCTACCTGGGTCGCGAGTC
419_FWD	ATGAAGCTTAGCTGTCCATTTCCGGATACCCATG
555_FWD	ATGAAGCTTAGCGTGTATATTCTTACGGTCCAGT
628_FWD	ATGAAGCTTAGCTCAGTTTGCATCCTTGTCCTGG
629_FWD	ATGAAGCTTAGCAAAGTAGTTTATCGTCGCGCAG
159_REV	GTAGTCGACCAATGCCTCGAAGGCCCCAAAGGTC
292_REV	GTAGTCGACCAATGCATCTCGCGTACGCCTGTGG
419_REV	GTAGTCGACCAATGCACACACCCAGAAGACGTGC
555_REV	GTAGTCGACCAATGCCAGCGTGTTAGCCCGACGC
628_REV	GTAGTCGACCAATGCCCGAGAGGGCTTGCGCTCT
629_REV	GTAGTCGACCAATGCCGAGGTTACACAGACACTG
159_stop_F	ATGAAGCTTTAGGTTGGGTAACCGG
292_stop_F	ATGAAGCTTTAGGCGGCTTAATGGGT
419_stop_F	ATGAAGCTTTAGTGTCCATAACCGGATACCC
555_stop_F	ATGGAAGCTTTAGGTGTATTAACTTACGGTCCAGT
628_stop_F	ATGAAGCTTTAGTCAGTTTAAATCCTTGTCCTGG
629_stop_F	ATGAAGCTTTAGAAAGTATAATATCGTCGCGC
159_R	GTAGTCGACCAATGCCTCGAAGG
292_R	GTAGTCGACCAATGCATCTCGCGTA
419_R	GTAGTCGACCAATGCACACACCC

555_R	GTAGTCGACCAATGCCAGCGTG
628_R	GTAGTCGACCAATGCCCGAGAGG
629_R	GTAGTCGACCAATGCCGAGGTTACA
e3_88058_F	ATGAAGCTTAGCCCCGTCTCCTGGATTCACGGTG
e3_88058_Stop_F	ATGAAGCTTAGCTGAGTCTCCTGGATTCACGGTG
e3_88058_R	GTAGTCGACCAATGCATAGCTTACCCCAGGC
e1_2234_F	ATGAAGCTTAGCCGCGGTATTCACCTAGGTCGGA
e1_2234_Stop_F	ATGAAGCTTAGCTGAGGTATTCACCTAGGTCGGA
e1_2234_R	GTAGTCGACCAATGCGTCCAAAACCCAGTGT
e2_1449_F	ATGAAGCTTAGCTACTGGAATAGCTCTATGGCGT
e2_1449_Stop_F	ATGAAGCTTAGCTGATGGAATAGCTCTATGGCGT
e2_1449_R	GTAGTCGACCAATGCGTCGGTATCAAACCGT
Outer_pFLAG_F	GCATAATTCGTGTCGCTCAA
Outer_pFLAG_R	AAAAGGGAATAAGGGCGACA
HFC-posA-stop-new-F	ATGAAGCTTAGCGTCTAGCGTCC
HFC-posC-stop-new-F	ATGAAGCTTAGCTAGGGTGGCC
HFC-common-posAC-stop-new-R	CTTGTCGTCATCGTCCTTGTAGTC
Neutral-oligo-F	GCATGAAGCTTAGC
Neutral-oligo-R	CTTGTCGTCATCGTC
Primers for the pET45b (+) vector inserts	
F4-pET-F	GATGAGCCCCGTCTCCTGGATT
F4-pET-R	CGACCTACAATGCATAGCTTACC
F32-pET-F	CCGATGAGCTACTGGAAT

F32-pET-R	ACCTACAATGCGTCGGTA
F600T-pET-F	GATGAGCCGCGGTATTCACCTA
F600T-pET-R	CCTACAATGCGTCCAAAACC
P159-pET-F	ATGAGCGTTGGGAAACCG
P159-pET-R	TCGACTTACAATGCCTCGAA
P419-pET-F	GATCCGATGAGCTGTCCATT
P419-pET-R	TTACAATGCACACACCCAG
P628-pET-F	ATCCGATGAGCTCAGTTTGC
P628-pET-R	GACTTACAATGCCCGAGAGG
Common-pET-F	GCGGATAACAATTCCCCTCT
Common-pET-R	ACCCCTCAAGACCCGTTTAG
Outer_pET_F	CACTTTTTCCCGCGTTTTC
Outer_pET_R	CGCCAATCCGGATATAGTTC

Table 2.5. Names of primers and their sequences used.

The names of primers containing ‘common’ or ‘outer’ in them are the primers flanking all the inserts and have their binding site in the vector backbone. In-frame stop codons depicted in red (underlined) were designed as controls for peptide expression upon IPTG induction.

2.2. Methods

This section describes all the methods used in this study.

2.2.1. General culture revival strategy

Cultures frozen (stored at -70°C) in glycerol (50%) stocks were used to streak on agar plates with appropriate media and antibiotics. The vials used to streak were never allowed to thaw to avoid repeated freeze-thaw cycles of the frozen cells. Cultures were scraped off from the top of the frozen vials under sterile conditions and streaked out to obtain maximum isolated colonies. The streaked plates were incubated overnight for 20-22 hours at the desired temperature before using single colonies to start overnight cultures on the subsequent day. Overnight cultures were started using single colonies and allowed to grow overnight for 16-18 hours at appropriate temperature at 250 rpm shaking conditions. This was the general strategy followed before starting any experiment unless otherwise mentioned.

2.2.2. Plasmid and genomic DNA extraction

2.2.2.1. Plasmid DNA extraction

Plasmid extraction was performed using the standard QIAprep® Spin Miniprep Kit (Qiagen) following the instructions provided. The final elution was done in 30 μL volume of the provided buffer in order to maximize the concentration of the plasmid DNA. Pure plasmid DNA samples were stored at -20°C indefinitely.

2.2.2.2. Genomic DNA extraction

Genomic DNA was extracted using the Sigma GenElute bacterial genomic DNA kit (Catalogue No. NA2120), following the exact protocol for the gram negative bacterial DNA extraction. Final DNA was eluted in 30 µL of the provided elution solution. Eluted DNA was stored at -20°C indefinitely.

2.2.3. Agarose gel electrophoresis

Generally, 1.5% agarose gels (SeaKem® LE Agarose) were made in 1X Tris-acetate EDTA buffer (Roth Rotiphorese® 50X TAE stock). Gels were loaded in Bio-Rad tanks (containing 1X TAE buffer) and run at 85-100 V for 45-60 minutes. Products were imaged using Bio-Rad molecular imager Gel Doc™ XR imaging system.

2.2.4. Cloning and transformation methods

2.2.4.1. Restriction digestion

Restriction digestion was performed using high fidelity HindIII-HF™ and SalI-HF™ sticky ended enzymes from NEB®. Double digestion was set up using 10 units (1 µL) of each of the enzymes in 1 µg of DNA, together with 5 µL of 10X CutSmart® buffer in a total volume of 50 µL made up by nuclease free water. The reaction was incubated for 1 hour at 37°C. The vector and insert DNA were digested individually before using them for ligation.

2.2.4.2. Ligation

Ligation was set up at room temperature for 10 m as described by NEB® ligation protocol. Ligation reaction was set up in a 1.5 µL microfuge tube on ice with 2 µL 10X ligation buffer, vector DNA (4.6 kb to insert fragment ratio was kept 1:3 (calculated using NEBioCalculator™ v1.10.1, formula: required mass insert (g) = desired insert/vector molar ratio x mass of vector (g) x ratio of insert to vector lengths), 1 µL of T4 DNA ligase and volume made up to 20 µL with nuclease free water. Ligated products were either immediately used for chemical transformation or stored at -20°C for a short time (1-2 weeks).

2.2.4.3. Chemically competent cells preparation and transformation

Competent cells were prepared by growing cultures overnight in 4 mL LB medium and inoculating 500 µl of the pre-culture into fresh 200 mL LB medium and allowed to grow at 37°C, 250 rpm until an OD600 of 0.45-0.55 was reached. The culture was collected in four 50 mL Falcon™ tubes and centrifuged for 10 m at 4°C, 3000 rpm. Pellets were gently resuspended in 1 mL chilled TBF-I solution (30 mM KOAc, 100 mM RbCl, 50 mM $MnCl_2$ and 10 mM $CaCl_2$) and filled up to 15 mL with the same. Tubes were incubated on ice for 1 hour followed by centrifugation for 10 m at 4°C, 3000 rpm. Pellets were then resuspended in 4 mL TBF-II solution (10 mM MOPS, 10 mM RbCl, 75 mM $CaCl_2$ and 15% Glycerol). Competent cells were ready. Cells were aliquoted (50 µL per tube) chilled and stored at -80°C up to 1 year for later use. For transformation, 50 µL of competent cells were thawed on ice and 100 ng (~ 5 µL) DNA was added gently. The mixture was incubated on ice for 30 minutes followed by a heat shock at 42°C for 60 seconds. Cells were immediately transferred on ice for 2-5 minutes. Pre-warmed 950 µL SOC medium (20 g bacto-tryptone, 5 g bacto-yeast extract, 0.5 g NaCl dissolved in 950 mL followed by addition of 10 mL of 250 mM KCL solution. Autoclaved and added sterile 5 mL 2 M $MgCl_2$ solution and 20 mL of 1 M sterile solution of glucose) was added to the cells and the cells were allowed to grow at 37°C with vigorous shaking for 1 hour (using a dry bath shaker). Transformants were plated on LB agar plates with 50 µg/mL ampicillin. Following controls were generally plated together with test for transformation: ligation control, competent cells control, cut vector only control and re-ligation vector control (without ligase). Plates were incubated overnight at 37°C and colonies were picked next day for confirmation using Sanger sequencing.

2.2.5. Sanger sequencing

Pure amplicons not more than 1 kb were used for Sanger sequencing. Common outer primers (see 2.1.7) were used to amplify all inserts. Forward and reverse primer reactions were carried out separately. DNA concentration of about 100 ng was used. For Sanger PCR, 1 µL of DNA, 0.5 µL of BigDye® terminator, 2 µL of 5X BigDye® buffer and 0.5 µL of 10 mM forward or reverse primer was used. A final reaction volume of 10 µL was made up using nuclease free water. PCR (of about 2.5 hours) was performed using the following parameters: denaturation at 96°C for 1 minute, followed by 30 cycles of- final denaturation, annealing and extension at 96°C for 10 seconds, 56°C for 15 seconds and 60°C for 4 minutes respectively. Tubes were then stored at 4°C indefinitely. PCR clean-up to remove the unincorporated BigDye® was done using BigDye XTerminator® purification kit (Thermo Fisher Scientific, Catalogue No. 4376484) using the standard kit procedure. Pure fragments were submitted for Sanger sequencing to in-house MPI Ploen sequencing unit. Output sequence files were analysed using Geneious R11 (version 11.0.5) or Geneious prime (version 2019.1.3) and CodonCode Aligner (version 7.0.1).

2.2.6. Candidate sequence purification from the library

Selected candidate sequences were pulled out from the initial random library. Specific primers were designed for each sequence of interest (list provided in Table 2.5) which were later used to amplify from the stored library plasmid DNA. Sequences were amplified using 2-step Phusion® high-fidelity PCR kit with the following PCR conditions: initial denaturation at 98°C for 30 seconds, followed by 30 cycles of – final denaturation for 98°C for 10 seconds with annealing at 72°C for 20 seconds. Final extension was performed at 72°C for 10 minutes followed by indefinite cooling at 8-12°C. Phusion® PCR kit uses a high-fidelity Phusion polymerase (with 5’→3’ polymerase activity and 3’→5’ exonuclease activity) which comes as a 2X master mix (kit instruction was followed for reagent volumes). The amplified products were purified using QIAquick PCR purification kit (from Qiagen) which were then used for downstream cloning. The purified amplicons and purified vector DNA were digested using HindIII-HF™ and SalI-HF™ restriction enzymes for 1 hour at 37°C (section 2.2.4.1) followed by purification using QIAquick PCR purification kit. Purified products were ligated with 1 µL T4 DNA ligase (protocol as per NEB® instructions) for 10 minutes at room temperature (bench top) using a 3:1 insert to vector ratio. Ligation products were transformed in already competent cells background *E. coli* strains using chemical transformation (section 2.2.4.3). Commercial competent cells were used for the *E. coli* K-12 DH10B strain (NEB® 10-beta high efficiency) but for other background strains, viz., *E. coli* K-12 MG1655, *E. coli* B REL606 and *E. coli* B REL607, competence was induced using TBFI/TBFII method described in section 2.2.4.3. Several transformation positive clones were freshly inoculated in 4 mL LB+ Amp media for preparing glycerol stocks next day and the same colonies were also used for colony PCR to confirm the presence of insert. Colony PCR was done by taking a part of a fully-grown colony from transformation positive plate and resuspending in 10 µL of sterile water. The suspension was heated at 98°C for 10 minutes and used as DNA for subsequent 2-step Phusion PCR (described before) using specific primers. Amplicons were sent for further confirmation by Sanger sequencing (section 2.2.5) only if they showed a band corresponding to the insert (195 bp) in 1.5 % agarose gel. Agarose gel electrophoresis was performed at 85-100 V for 30-45 minutes.

2.2.7. Growth assays

2.2.7.1. Growth measurements using the plate reader

All growth curves were performed on Tecan M nano+ plate reader. Cultures from frozen stocks were streaked on appropriate media plates with ampicillin (50 µg/mL) and incubated overnight at the desired temperature (usually 37°C). Next day single colonies were inoculated in 200 µL of desired medium with ampicillin in 96 welled plates and incubated at desired temperature (usually 37°C) with vigorous shaking for 16-18 hours. Following day, growth curves were set up by adding 4 µL of overnight culture to 200 µL of medium with ampicillin with or without 1 mM IPTG for induction of expression. Controls did not contain IPTG. Usually more that 3 replicates were used (every replicate came from single colony). Growth curves were performed for 24 or 48 hours with 5 minutes of orbital shaking before reading OD600 every 10 minutes at the desired temperature. Note: Some of the initial experiments were performed with a reading interval of 30 minutes for 24-48 hours with continuous shaking (mentioned wherever relevant). This was changed later to a shorter interval to obtain a finer growth curve. Initial growth curves were performed by Ellen McConnell, where absorbance readings were taken every 30 minutes for 24 hours with continuous orbital shaking in BioTek® Epoch microplate reader (mentioned wherever applicable).

2.2.7.2. Manual growth assessment using colony forming unit counts

Manual time series experiment was performed to obtain the colony forming units in growing cultures. OD600 measurements in the plate readers are a measure of turbidity and could be overestimated by factors like exopolysaccharide production. Strains were revived by standard procedures and growth measurements were started in larger volumes of 5 mL with a starting dilution of 1:100. After every hour of growth with or without IPTG induction at 37°C, 250 rpm, cultures were plated on LA+Amp plates at an appropriate dilution and incubated overnight at 37°C. Colonies were counted using a colony counter and CFU. mL^{-1} at each time point was calculated. Growth curves were performed for 13 hours.

2.2.8. Frequency of emergence of suppressors

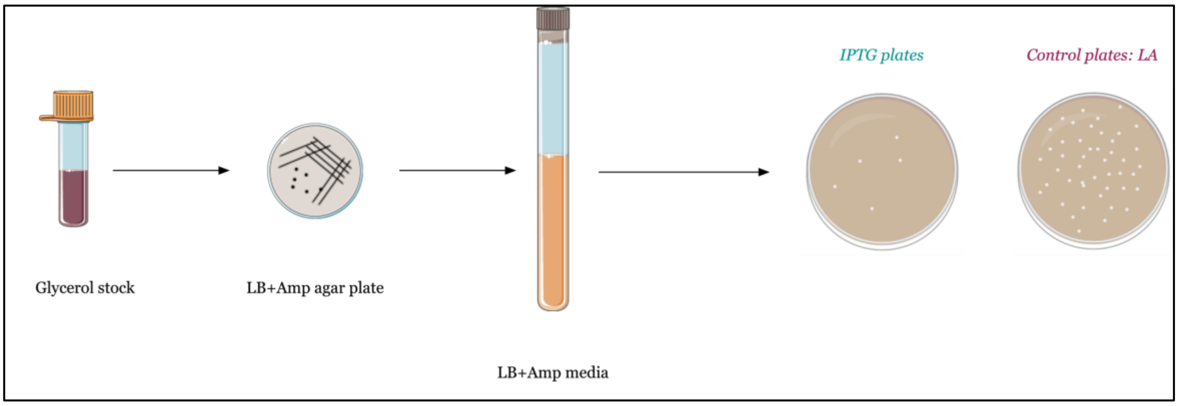

Figure 2.2. Experimental design for screening of suppressor clones.

Cultures were streaked from glycerol stocks on an agar plate with appropriate media and grown overnight at 37°C. Next day, single colony was inoculated per replicate tube containing 4 mL appropriate media and incubated at 37°C with 250 RPM for 16-18 hours.

> The overnight culture was then used (at an appropriate dilution) to spread on a solid media plate with and without IPTG and further incubated at 37°C until visible colonies appeared.

Suppressors were the evolved clones that did not show fitness defect upon induction (unlike corresponding parents that had 5-10 hours of growth lag). Appropriate candidate strains were streaked on LB+Amp agar plates and allowed to grow overnight at 37°C (Figure 2.2). Single colonies were inoculated in 4 mL LB+Amp media and incubated for 16-18 hours under shaking conditions (250 rpm) at 37°C. On the subsequent day, overnight cultures were plated on LA+Amp agar plates containing 1 mM IPTG. This means that the peptide production in the respective clones is induced when cultures are plated on IPTG agar plates. Frequency is calculated by taking the ratio of total number of colonies of IPTG agar plates and the total number of colonies on uninduced LA+Amp agar plates (expression control).

2.2.9. Competition assays

Competitive fitness assays were performed using two strain backgrounds viz., *E. coli* B REL606 and *E. coli* B REL607 using the red-white selection as described (Lenski et al., 1991).

Strains REL607 and REL606 are ancestors, with a single nucleotide difference in the arabinose utilization gene of REL607, such that it has acquired the ability to metabolize arabinose. The mutation in this strain was spontaneous and is neutral in this strain making it suitable for red-white selection. The two competitors can be distinguished by their arabinose utilization phenotypes; (REL606) Ara$^-$ and (REL607) Ara$^+$ strains that produce red and white colonies respectively on TA agar plates.

The two backgrounds were transformed with desired candidates (see section 2.2.4.3). Desired candidate strains were streaked on M9+Glucose+Amp agar plates (see section 2.1.3) and incubated at desired temperature (33, 37 or 40°C) to obtain single colonies. 5-8 colonies were inoculated the next day in 4 mL

M9+Glucose+Amp media per competitor strains and allowed to grow at desired temperature (33, 37 or 40°C) for 18 hours with shaking at 250 rpm. On the subsequent day, 1:1 volume of overnight cultures was used from each competitor strains and mixed thoroughly in a 96-welled plate (50 μL each strain). After mixing them in 1:1 ratio, 4 μL of the mixture was inoculated in a fresh M9+Glucose+Amp media (with or without IPTG) and incubated for 24 hours at the desired temperature with shaking (250 rpm). The strains in tubes compete for resources over 24 hours. Simultaneously, the mixture was plated (T_0) on TA+Amp plates (see section 2.1.3 for TA plate preparation) at an appropriate dilution such that the colony count is between 30-200 (for statistical significance range) and incubated at 37°C overnight. Cultures were taken out after 24 hours (T_{24}) and plated at an appropriate dilution and incubated at 37°C overnight. Colonies from the previous day were counted where a 50-50 ratio was expected. Next day, colonies from T_{24} were counted and relative fitness of strains was determined.

2.2.9.1. Calculating relative fitness of strains

Relative fitness is calculated using the following formula: (as described in (Lenski et al., 1991))

$$w = {ln\left(\frac{A_f}{A_i}\right)} \Big/ {ln\left(\frac{B_f}{B_i}\right)}$$

where w is fitness, A and B are the two strains colony forming units per millilitre, i and f are initial and final time points (T_0 and T_{24} here); ln is natural logarithm. Individual terms in the numerator and denominator are known as Malthusian parameters (m), which reflect the change in population density of the particular strain (here A or B) over time.

Relative fitness values were used to determine fitness of a particular candidate and a swapped control was always tested where the backgrounds were swapped and relative fitness was measured. This was easily possible because all candidates exist on a vector that could be chemically transformed in any strain background.

2.2.10. Western blotting and detection

Western blots were performed to test presence of protein expression after induction with 1 mM IPTG. Samples were grown overnight usually at 37°C at 250 rpm unless otherwise mentioned. Overnight cultures were then inoculated next morning into a fresh 4-5 mL LB+ Amp media and allowed to grow up to an OD of 0.4-0.5. About 1.4 mL of uninduced cultures were aliquoted at this point for the no expression control. Samples were then induced with 1 mM IPTG and were allowed to grow for 1 hour. An aliquot of 1.4 mL was taken. All samples were spun down at maximum speed for 3 minutes and supernatant was discarded. Pellets were resuspended in 30 µL of Laemmli sample buffer with 5% β-mercaptoethanol. Samples were incubated at 98°C for 5 minutes. A volume of 10 µL of each sample was loaded onto a 4-20% tris-glycine gel (Bio-Rad) and allowed to run for 2 hours at 70 V. The proteins were then transferred to polyvinylidene difluoride membrane (PVDF) for 15 min at 13 V using a semi-dry electro blot transfer unit (Bio-Rad). The membrane was washed 3 × 10 minutes in tris buffered saline (TBS; 150 mM sodium chloride, 50 mM tris-HCl, pH 7.6) with 0.1% tween-20 (TBST) and then blocked in 5% powdered milk (1% fat) dissolved in TBST with gentle shaking at room temperature for 1 hour. Monoclonal mouse anti-FLAG M2 antibody was added (diluted 1 in 2000 in 2.5% milk PBST). The membrane was incubated overnight with shaking in a cold room (4 °C). Next day the membrane was washed 3 × 10 minutes in PBST with shaking. Secondary goat anti-mouse alkaline phosphate (section 2.1.5) was added (diluted 1:5000 in 2.5% milk TBST) and incubated for 1 hour at room temperature with shaking. The membrane was then washed 3 × 10 minutes in TBST on a shaking platform. Colorimetric detection was performed using combination of NBT (nitro-blue tetrazolium chloride) and BCIP (5-bromo-4-chloro-3'-indolyphosphate p-toluidine salt) solutions which yields an intense, insoluble black-purple precipitate when reacted with alkaline phosphatase. 45 µL NBT solution and 35 µL BCIP solution were added to 10 mL of 1X DIG buffer solution. Membrane was flooded with the above solution and stored in dark colour development. Normal light imaging was performed with Bio-Rad molecular imager

Gel Doc™ XR imaging system.

2.2.11. Antibiotic sensitivity test: spot dilution assay

LB agar plates were made with variable concentrations of ampicillin (i.e. 50, 75, 100, 125, 150, 200 and 250 µg/mL) in order to test antibiotic sensitivity of suppressor strains compared with the respective parent. Desired strains were grown by streaking from glycerol stocks and allowing to grow at 37°C overnight on LB+Amp (50 µg/mL) agar plates. On the subsequent day, single colonies were inoculated into 200 µL LB+Amp media in 96-welled plates and allowed to grow overnight under shaking conditions (250 rpm) at 37°C. Ability of the strains to grow on increasing concentrations of antibiotic was tested by making a serial dilution of fully grown cultures (up to 10^{-7}) and then making a 10 µL spot using multichannel pipette on previously prepared agar plates containing variable antibiotic concentration. The minimum inhibitory concentration was estimated by noting the concentrations from where specific strains were inhibited due to high antibiotic concentrations (no turbid spots) compared to respective parent strains.

2.2.12. Microarray

2.2.12.1. Sampling of bacterial cultures at different time points

Agar plates with appropriate media were streaked for different strain samples to be tested, were incubated at appropriate temperatures (37°C or 40°C) under static conditions. Different media (LA+Amp or M9+Amp) and temperatures were used but kept constant throughout the experiment. Similar culture and temperature conditions were used in order to acclimatize the strains to particular environment in which they were being tested. After about 20-22 hours, single colonies (in 3-5 replicates) from the streaked plates were inoculated into 5 mL of liquid media (same as the one used in agar plates) followed by incubation at appropriate temperatures with shaking at 250 rpm overnight (16-18 hours). Inoculated tubes were allowed to grow until they reached the exponential phase which was determined by measuring the optical density (OD_{600}). An OD_{600} of 0.4-0.5 (standardized prior to experiment) was considered adequate for cultures to be in exponential phase. At this point, 500 µL of cultures were aliquoted and 1 mM IPTG was added to the remaining cultures (4.5 mL). Aliquoted cultures were centrifuged for 5 minutes using a table-top mini centrifuge at maximum speed (13000 rpm) and pellets were stored at -20°C after treatment with RNAprotect reagent (Qiagen, section 2.2.12.2). This was T_0. The remaining cultures, induced by adding 1 mM IPTG, were immediately moved back to the shaker following the first aliquot and allowed to grow further. T_1 aliquots were taken after 1 hour of letting cultures grow under induction and T_2 aliquots were taken 16 hours of growth post-induction. All time points were taken in similar manner as explained above (for T_0). Pellets frozen in RNAprotect were used for RNA extraction within one week.

2.2.12.2. RNAprotect treatment and total RNA extraction

Cultures aliquoted from different time points were vigorously (5 second vortex) resuspended in 2 volumes equivalent of RNAprotect bacteria reagent (for 500 µL cultures, 1 mL reagent was added, Qiagen). The mixtures were incubated for 5 minutes at room temperature and then centrifuged for 10 minutes at 5000 x g. Supernatants were decanted and pellets were stored at -20°C up to one week. RNAprotect bacteria reagent enters the cells and protects the RNA from denaturation.

Total RNA extraction was performed using the RNeasy® Mini kit (Qiagen, Catalogue no. 74106) following the kit protocol. Final elution was performed with 40 µL of pure water. Total RNA samples were stored in -70°C until further use.

2.2.12.3. Hybridization for microarray

Hybridization was performed in three steps described below, however the described protocol is a summary of full protocol from Low Input Quick Amp WT labelling kit - Manual part no. G4140-90042, Version 2.0, August 2015.

Part 1: Sample preparation

- One-Colour spike mix preparation
- cDNA synthesis
- Labelling
- Purification of the labelled/ amplified RNA
- cRNA quantification

Part 2: Hybridization

- Hybridization samples preparation
- Hybridization assembly preparation

Part 3: Microarray

- Microarray slides preparation
- Imaging and feature extraction

The labelling kit generated cyanine labelled cRNAs which were amplified using the WT kit primer mix (mixture of oligo dT and random nucleotide-based T7 promoter primers) generating cRNA from samples. The provided spike-in controls were also labelled and amplified with the samples. Labelling, hybridization, washing and scanning were performed using the standard protocol from Agilent user guide (Low Input Quick Amp WT labelling kit - Manual part no. G4140-90042, Version 2.0, August 2015). We performed one colour microarray using the already available Agilent *E. coli* microarrays 8 X 15K, P/N G4813A, using the design ID 020097, for complete gene probes list.

2.2.12.4. Microarray data analyses using Limma package from R

Limma package in R (version 3.6.1) was used for analysing microarray data. Data analysis step by step modified from documentation available in R package help:

- Microarray image analysis output files included IDs, names of probes and annotation information. Each array (i.e. 1 sample) produced one such text file. A ‘targets’ file was created which was a normal text file that had all information which was needed later during separating the samples. Targets file had information like file name, sample name, treatment, environment, etc. It could be read in R using a function called `readTargets()`.

- Raw data files (text format) were read into `read.maimages()` function.
- A background correction was done using `backgroundCorrect()`and normalization between arrays was performed using the function `normalizeBetweenArrays()`.
- `avereps()` function was used to average all replicate probes. Since the chip had around 15000 probes, many probes were in replicates which could be averaged before analyzing.
- Control probes, probes with no annotation and probes with high background noise were filtered out as they do not represent the expression of samples (this step was performed after the normalization and background correction).
- Different levels were assigned to the dataset (i.e. each level is each sample) of all arrays, which could later be used to make various comparisons.
- A design was made using the function `model.matrix()` that created a design matrix using input variables. The design matrix contained rows corresponding to arrays and columns corresponding to set levels (coefficients to be estimated).
- A linear model was then fitted for each gene in series of arrays using `lmFit()`. Input file was the filtered Elist generated from raw expression data from the array chip which had log-expression values (rows corresponded to genes and columns to samples). Design was provided with this where rows corresponded to arrays and columns to coefficients to be estimated.
- `makeContrasts()` function was used to make comparisons between different set of samples within and between experiments.
- After desired contrasts were selected, `contrasts.fit()` function computed estimated the coefficients and standard errors for all provided contrasts.

- Genes were ranked in order of differential expression and t-statistics, moderated f-statistic and log-odds of differential expression were computed using Bayes modification of standard errors to a common value using the function `eBayes()`.

A summary table was generated together with the function `decideTests()`, which identifies all genes which were significantly differentially expressed for each contrast from a 'fit' object which was the output of linear fitted expression data from step 9.

- A Venn diagram of the differentially expressed genes within contrasts provided could be plotted using `VennDiagram()` function.
- A log fold change versus log expression could be plotted using function `volcanoplot()`.

2.2.12.5. Gene ontology analysis using ShinyGO v0.61 graphical tool

Gene ontology was used to be able to visualize differences between different samples w.r.t general biochemical and regulatory pathways. In general, GO term analysis was performed using top 100 differentially expressed gene-list that arose as a result of any comparison between control and test (see 2.2.12.1 for experimental design and sampling). ShinyGO v0.61 was used to output interaction networks and tables (Ge et al., 2019). Note that gene ontology was used only for general visualization and no inferences were made based on it.

2.2.13. Whole genome re-sequencing

Whole genome re-sequencing was performed with 6 (+ 1 control) parent genomes and 18 (+ 3 corresponding controls) suppressor genomes (3 per parent). Two different genomic backgrounds were sequenced; *E. coli* K-12 DH10B and *E. coli* K-12 MG1655. Here, controls were strains containing empty vector. DNA samples were extracted (section 2.2.2.2) from overnight grown cultures as described in section 2.1.3. All samples were sequenced using in-house sequencing facility at the MPI, Ploen.

2.2.13.1. WGS parameters

Whole genome re-sequencing was performed using the Illumina NextSeq® 550 sequencing platform. The multiplexed library was prepared using the Nextera DNA Flex Library Prep Kit followed by paired-end sequencing with average read length of 2 x 150 bp using NextSeq 500/550 v2.5 reagent kit. Number of paired reads that passed the filter were more than 4 million per sample, which provided about 40X read depth. Whole genome re-sequencing was performed at the Max Planck Institute for Evolutionary Biology, Ploen, Germany by our in-house sequencing team, headed by Dr. Sven Künzel.

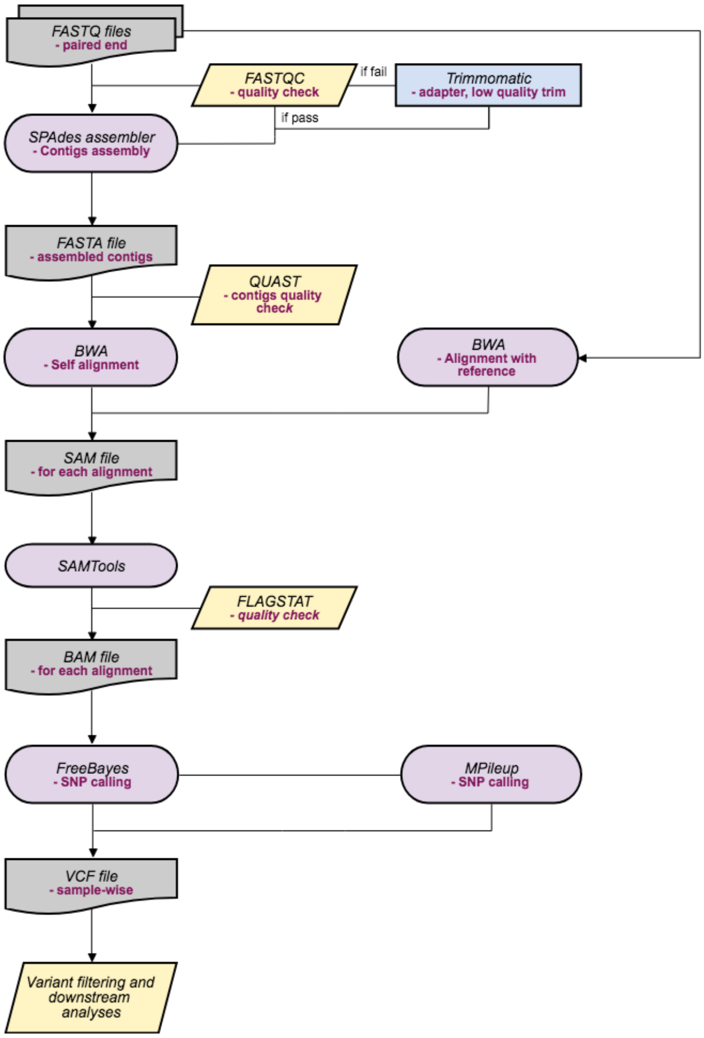

Figure 2.3. Pipeline for processing paired-end reads of the whole genome resequenced samples.

The above pipeline was used to process all paired-end reads for every sample. Variant calling was also performed separately using Geneious Prime.

2.2.13.2. Polymorphism analyses

Analyses were performed using several platforms and methods to reconfirm the results. Whole genomes were assembled de novo and against the available reference genomes downloaded from NCBI database. All reads were also assembled using an open source software called *Breseq* which is specifically designed to analyse *E. coli* clonal population data (Figure 2.4, (Deatherage and Barrick, 2014)). Different methods were employed for accuracy in determining polymorphisms and structural variations. The output from *Breseq* was compared with the output analysed separately using assemblers and variant callers from Geneious R11 (version 11.0.5) or Geneious prime (version 2019.1.3) after setting thresholds manually. Custom genomes were built using `gdtools` function in *breseq*, where all polymorphisms in the laboratory ancestors were applied to create a modified lab reference genome. This allowed automatic filtering of mutations that were irrelevant due to their presence in the lab strains. Note that it is quite common to have some new mutations in the lab ancestor compared to the NCBI reference due to propagation of strains through labs (and cycles of freeze thawing), making it absolutely crucial to have lab ancestor strains resequenced.

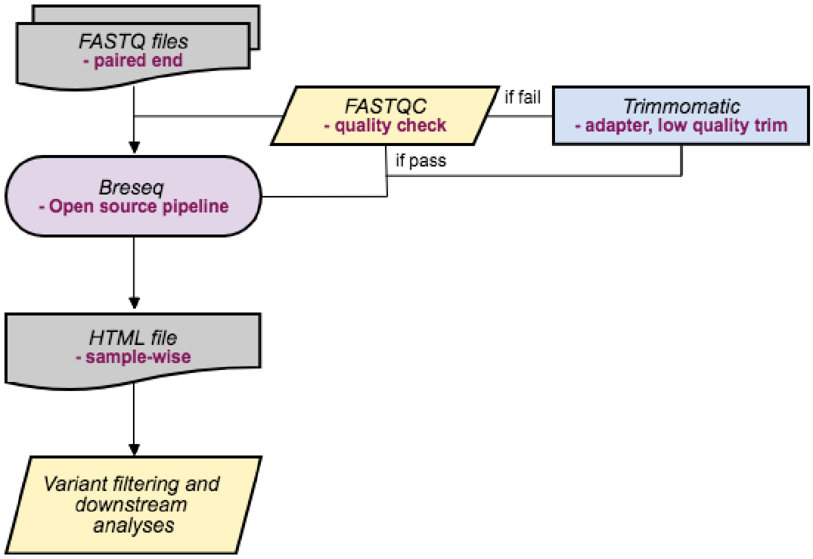

Figure 2.4. Pipeline using an open source tool *Breseq*. The open source software called *Breseq* was used to analyse all samples. This platform was specifically designed by authors to analyse clonal *E. coli* datasets and hence was used for this study.

Chapter 3. Results I

Effects of expressing negative random coding sequences in Escherichia coli

1.1. Introduction

Organisms utilize a limited set of proteins even though there is a vast number of combinatorial possibilities for a string of amino acids (20^n, where n is the length of the polypeptide) to produce a protein (Chiarabelli et al., 2006). It is therefore possible to search for novel functions in random stretches of polypeptides that may never have been explored in nature. Random stretches of sequences may provide the raw material necessary for evolution of new functional peptides. The mechanism that leads to evolution of new genes and proteins without any making use of any previously existing protein coding sequence (for example in case of duplication) is known as *de novo* evolution (Tautz, 2014). Previously in our lab, a random (coding) library cloned into an inducible expression vector was studied to look for bioactivity after expression in a model bacterium *Escherichia coli* (Neme et al., 2017). Most (about 75%) of the random sequences showed decreased abundance over time, i.e. they declined in frequency when compared to the starting frequency (see introduction Figure 1.3A). This decrease in frequency was considered to be a consequence of a fitness decrease in the bacteria that were expressing a given random peptide in question. Millions of unique random peptide variants were present in their experimental *E. coli* population at the same time. Random sequences have a negligible probability of matching to a known biological sequence but they might interact among themselves (e.g. due to aggregation) or with the cellular components by spontaneous interactions. This may also have played a role in altering the frequencies of sequences at the end of the time series experiment that the authors studied. Hence, the following step was to study the effects of individual random sequences on host fitness.

It can be argued that most heterologous peptides over-expressing inside a bacterial host are deleterious, simply due to the cost of producing large amounts of unwanted proteins (Weisman and Eddy, 2017). The host is thought to be battling with: translational demand, non-optimal codon usage, protein aggregation, etc., while attempting to reach an optimum growth in the available nutrient conditions (Bolognesi and Lehner, 2018). Contrary to the idea that random sequences are mostly harmful to cells producing them, is a study where authors have shown that random sequences are generally well-tolerated *in vivo* (Tretyachenko et al., 2017). The authors show that highly disordered random sequences have a low propensity to aggregate within the cells. Although biological sequences have been shown to reduce fitness of the host upon over-expression (Kafri et al., 2016), random sequences have an inherent property of high intrinsic disorder, which allows them to be well-tolerated inside hosts. Different random sequences would have different degrees of effects on the host and could potentially trigger wide range of responses followed by adaptation especially to the more harmful sequences. The effects of individual random sequences on the host have not been studied enough in the light of *de novo* gene evolution although it makes intuitive sense to do so.

This chapter describes in details the effects of expressing individual candidate random peptides on the bacterial host, *Escherichia coli* sp. It specifically deals with negative spectrum sequences (NEG_Peps; Figure 1.3A, red circles), which have shown a reduction in growth rate without directly killing the cells (Neme et al., 2017). Due to the milder deleterious effects i.e., not lethal, specific interaction effects could be predicted. Each candidate was ligated into an expression vector, which enabled understanding of direct effects of individual random peptides on the fitness of bacteria. Assuming the negative effect of a given random peptide, in sense that it has only one or few interactions inside the cells, would still constitute as a new genetic function encoded by the random sequence. Since the peptides in question are thought to be deleterious, it is also expected that *E. coli* would adapt to this stress by reducing the expression of harmful peptides or eliminating the target of the peptides altogether. Interestingly, for all the candidates it was easily possible to obtain suppressor of the mutant phenotypes. Through the characterization of these suppressors, it was hoped that possible interaction partners of the candidates could be identified. Whole genome re-sequencing of the evolved clones revealed the underlying genetics of suppressor evolution, all of which will be described in the chapter.

3.1. Aims

To understand effects expression of candidate random coding sequences on bacterial fitness, I selected six candidates that decreased in abundance over time in the previously reported serial passaging experiment (Neme et al., 2017). The effects of each of these six individual candidate sequences were studied in detail, with the following specific aims:

1. To assess the effect on growth of each candidate peptide in four different *E. coli* backgrounds (K-12 DH10B, K-12 MG1655, B REL606 and B REL607).
2. To understand underlying basis of any growth effects observed in aim 1.
3. To test differential mRNA expression profiles of strains expressing candidate peptides at different time points in order to gain a better understanding of overall response of bacteria to representative candidate peptides.
4. To isolate and characterize suppressor of mutant phenotype
5. To identify interaction partners of the candidates by whole genome re-sequencing approach

I address each of the above aims extensively in following sections, with the overarching aim of gaining insight into a broader question of how bacteria may respond to novel random peptides.

3.2. Findings

Six candidates were chosen from the previously described experiment as representatives of the negative fold-change (candidates that decreased in frequency) category from the random library as described in section 1.3. Only a handful sequences were chosen because the aim was to study the effects of individual sequences on bacterial fitness in high level of detail. The main goal was to understand whether random sequences can have biological functions through specific interactions with other proteins.

3.2.1. Selection of negative candidates from library

Deep sequencing data analysed in the previously mentioned study (Neme et al., 2017), allowed us to select interesting candidates on the basis of differential counts and statistical significance of their decrease. Six candidate sequences (full sequence list in Table 2.3) were selected from this data in order to explore their individual effects on bacterial fitness (Table 3.1). In an effort to achieve an overview of possible effects, I chose individual sequences based on the changes in their abundance (from high to low level fold changes) over time. Selected candidates were extracted from the library and transformed into four different *E. coli* backgrounds (using the same multicopy vector pFLAG-CTC) for further investigation. Unlike the random insert library expressed in *E. coli* sp. from the previous study, each of my strains express a unique random peptide of choice (methods 2.2.6).

Sequence name	Original ID	Log2 fold change	P-value
NEG_Pep1	PEPNR00000000159	-7.05	6.5E-23
NEG_Pep2	PEPNR00000000292	-6.06	2.2E-11
NEG_Pep3	PEPNR00000000419	-6.56	2.4E-11
NEG_Pep4	PEPNR00000000555	-5.44	3.9E-06
NEG_Pep5	PEPNR00000000628	-4.79	3.5E-04
NEG_Pep6	PEPNR00000000629	-2.85	8.2E-02

Table 3.1. Candidates selected on the basis of negative fold change and p-values. Candidates of different fold change values were picked to maximize the range of the representative sample. Sequences of candidates can be found in Table **2.3**. The values are extracted from the deep sequencing data tables published previously (Neme et al., 2017) in Dryad: http://dx.doi.org/10.5061/dryad.6f356**.**

The individual sequences in the library were designed with common adjoining regions that included all features necessary for the translation of the insert, producing random coding peptides upon induction (Figure 3.1). The coding feature was introduced to mimic a 'protogene', which is a partial gene with spontaneously evolved coding capabilities, as described in the life cycle of a gene illustration (Neme and Tautz, 2014). The insert and vector design was the same as it was in the previous study.

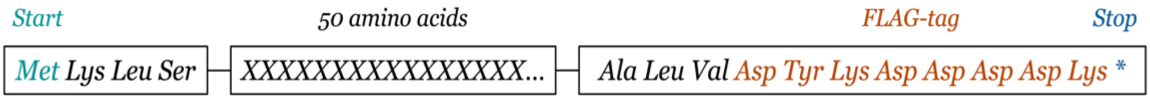

Figure 3.1. Design of random insert cloned into pFLAG-CTC vector.
The random sequence was designed with start and stop codons together with a FLAG tag. Cloning was done using flanking restriction sites (not shown here) which were also present in the plasmid multiple cloning site (MCS).

3.2.2. Negative candidates show a fitness disadvantage upon expression

The phenotypes of the strains expressing negative candidate sequences were investigated using growth curves. Growth curves were performed (using Tecan Infinite® M Nano$^+$ microplate reader) for 24 hours at 37°C with intervals of 30 minutes between absorbance readings at 600 nm. Orbital shaking was performed throughout and only shortly stopped before the absorbance was measured. *E. coli* K-12 DH10B strains expressing each of the six candidates (NEG_Pep1- 6) on the vector (pFLAG-CTC) are shown here. Upon induction of expression with 1mM IPTG, all candidate strains show a statistically significant lag in growth compared to the respective non-induced controls (see Figure 3.2A & B). After the prolonged lag phase, all strains are seen to have resumed normal growth rates (Figure 3.2C, p>0.05). Strains that harbour empty plasmids show no such differences in their growth patterns following induction (Figure 3.2, panels ‘pFLAG’).

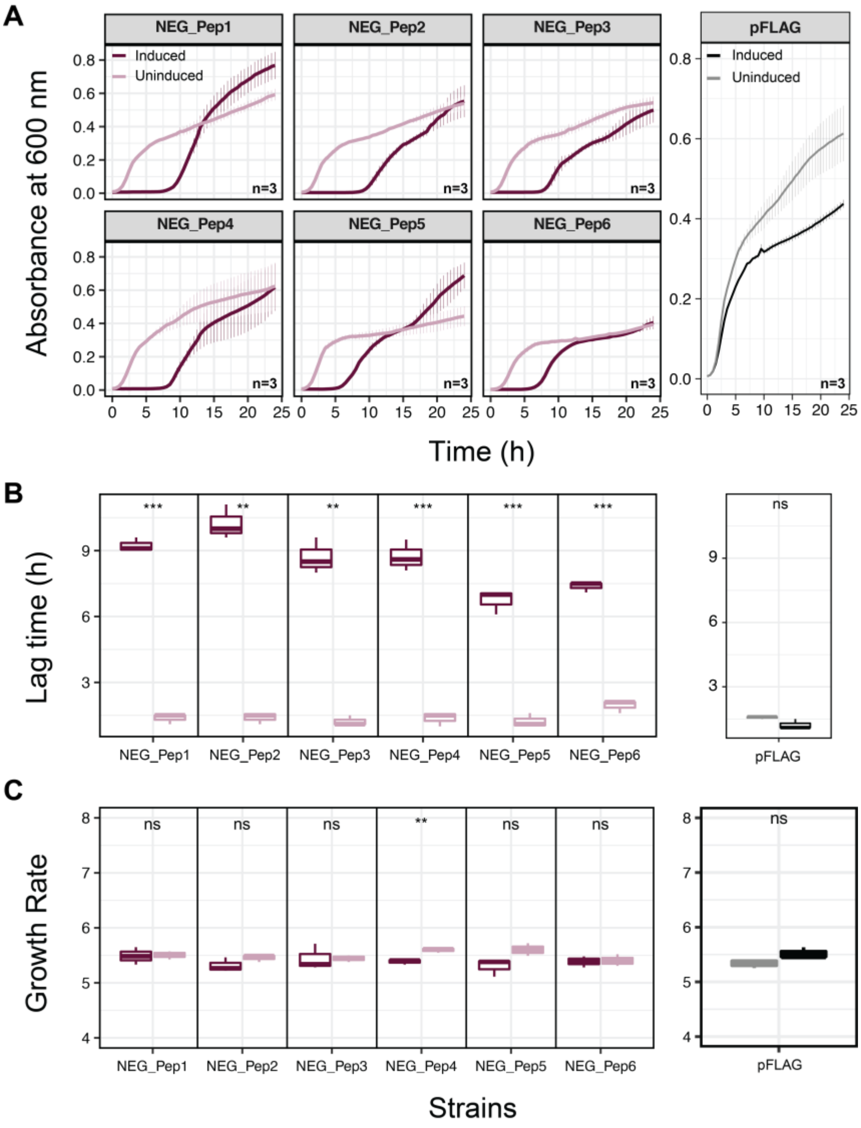

Figure 3.2. Expression of NEG_Pep1-6 leads to delayed growth in DH10B strains. A) Growth measurements of NEG_Peps and pFLAG control are shown. A prolonged lag is observed in NEG_Peps1-6 upon expression (maroon), whereas the corresponding non-induced strains (pink) show normal growth that is analogous to the pFLAG controls (black and grey). 'n' is the number of replicates. Error bars represent standard error of the mean (SEM). **B)** Lag time in hours shows a significant difference in the all induced NEG_Peps (1-6; maroon) compared to the controls (pink). pFLAG controls show no difference between induction (black) and no induction (grey). **C)** Exponential growth rates (calculated from sliding window of the consecutive values of the slope of the curves) show no significant differences (p-values > 0.05) in induced versus non-induced strains, except for NEG_Pep4 where induction causes decrease in the growth rate. Maroon colour are NEG_Peps after induction with 1 mM IPTG and pink colour shows the corresponding non-induced strains. Black is pFLAG after induction and grey is without induction. All candidates are in the inducible vector pFLAG-CTC against a *E. coli* K-12 DH10B strain background. Growth curves were performed at 37°C for 24 hours and readings were taken every 30 minutes (methods 2.1.3). For **B)** and **C)** statistics performed using Student's t-test. P-values: ** < 0.01, *** < 0.001 and ns > 0.05; non-significant. Lag time and growth rates were calculated using a command line program GrowthRates v4.2 (Hall et al., 2014).

All candidates are under the control of a strong promoter, which upon induction with IPTG, start expressing the peptides. The empty plasmids on the other hand, produce a 38 amino-acid long peptide due to the presence of a multiple cloning site after the promoter and before the stop codon. This or the presence of IPTG, did not affect the growth patterns of the cells, although small, statistically insignificant differences may still exist (but no visible lag observed). Overexpression of candidates causes a lag in the growth of the host, followed by a climb in the slope of the curve to compensate the delay. Given the prolonged lag phase, the host strains certainly become unfit upon candidate peptide overexpression, but they manage to cope with the deleterious peptides and in some way, restoring normal-level growth rates within 24 hours (similar to the controls). The maximum OD for some candidates shows a variable value, but this was ignored as at higher cell densities cells tend to clump and the OD_{600} readings are affected (for example, see Figure 3.2A- NEG_Pep1 and 5).

3.2.2.1. Prolonged lag remains across different strain backgrounds

The candidate NEG_Peps showed a distinct and reproducible pattern in their growth dynamics in the strain *E. coli* K-12 DH10B derivative (commercial name: NEB® 10-beta). I extended the phenotypic exploration of two candidates (NEG_Pep2 and 3) to other strain backgrounds for two key reasons:

- To test host range of NEG_Peps by testing them on different backgrounds.
- To test effects on wild type strains, since commercial DH10B derivative (or NEB® 10-beta) is a highly-modified strain with large number of deliberately introduced mutations for the ease of cloning and transformation (see strains 2.1.1).

To determine the host range of the deleterious effects displayed by NEG_Peps, each candidate was transformed into chemically competent cells (methods 2.2.4) of two different genetic backgrounds viz., *E. coli* K-12 MG1655 and *E. coli* B REL606 (2.1.1). Growth measurements were performed every 10 minutes for 24 hours in

LB medium at 37°C under shaking conditions (250 rpm) as described in 2.2.7.1. As described before in Figure 3.2, the expression of candidate peptides has a growth disadvantage in the DH10B background due to a prolonged lag in their growth. Candidates tested with the two new strain backgrounds showed an analogous growth lag, although the time of the lag reduced almost by half (Figure 3.3). Two candidate sequences (NEG_Pep2 and NEG_Pep3) tested against two genetic backgrounds (MG1655 and REL606) are shown. The candidates have a deleterious effect (growth defect) not only on the commercial DH10B strain derivative (original background used) but also on two distinct wild type strains studied here. The deleterious effect of random peptides is not limited to the host in which they are enriched but show a relatively broad range.

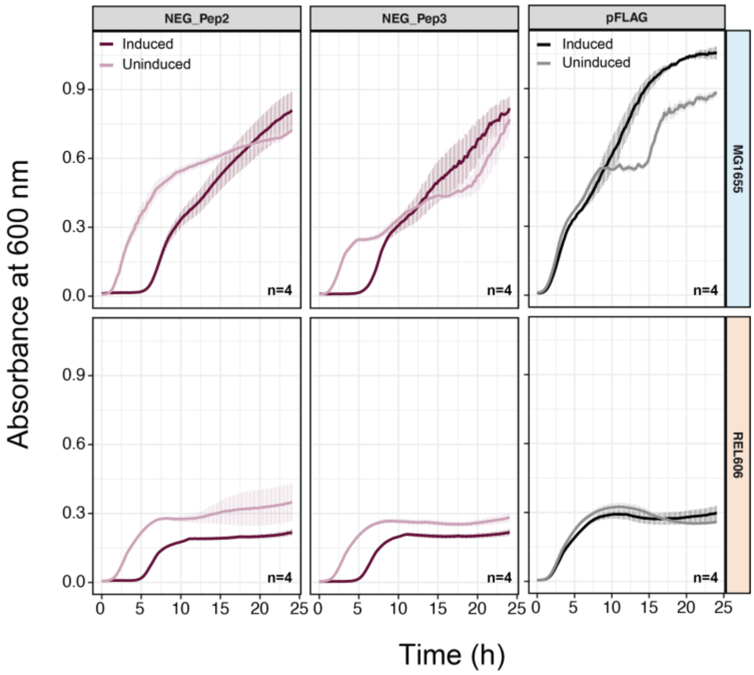

Figure 3.3. Deleterious effects of random peptides can be reproduced in different backgrounds. Growth curves of NEG_Pep2 and NEG_Pep3 on two genetic backgrounds, viz. K-12 MG1655 and B REL606. Induced NEG_Peps (maroon) show a prolonged growth compared to corresponding non-induced controls (pink). Induction was done with 1 mM IPTG. pFLAG controls are represented in black when induced and grey when non-induced. Growth measurements were performed at 37°C for 24 hours with OD reading taken every 10 minutes with 5 minutes shaking prior to OD reading (2.2.7.1). The rows correspond to two strain backgrounds: K-12 MG1655 (top) and B REL606 (bottom). Error bars represent standard error of the mean (SEM). ‘n’= no. of replicates.

Note that the maximum absorbance values are different for different strains. MG1655 strains can grow up to an OD_{600} value of about 1.2 whereas REL606 strains can only grow up to an OD_{600} of 0.5. These are the inherent growth abilities of the respective strain backgrounds on LB medium at 37°C and have nothing to do with candidate sequences *per se*. For example; the empty vector strain in MG1655 background shows a distinct growth pattern (2-step dynamics) only in the non-induced state. This may possibly be due to the diauxic growth phenomenon where the cells utilize simple carbon sources like glucose first and then after a small lag (for acclimatization) switch to the complex carbon sources like lactose. Since LB is a rich medium a similar mechanism might be at play. A comparative growth assay with pFLAG was always performed as control where strains simply maintain the plasmids (due to antibiotic selection in media) during growth. The deleterious effects of NEG_Peps (at least for NEG_Pep2 and 3) are not restricted to the host they are enriched in (i.e. DH10B), but are reasonably broad.

3.2.2.2. Introduction of in-frame stop codons can rescue the deleterious phenotype

In theory, engineering an in-frame stop codon at the beginning of the sequence, should rescue the fitness defect caused by the harmful peptides, because a truncated peptide would be created. If the fitness defect is indeed caused by peptides, truncated products may not have the disadvantage within cells. To this end, I engineered in-frame stop codons in all six candidates using standard high fidelity PCR (Figure 3.4). Stop codons were placed in the initial part of the sequences (truncated products contain only three to six amino acids) such that only a small product is formed, i.e. most of the original coding sequence is removed. Specific forward primers were designed by changing one or two nucleotides to create in-frame stop codons, which incorporate by annealing and amplify the mutated sequence by PCR. Although the truncated peptides are still under a strong promoter, no accumulation of full-length proteins is expected.

NEG_Pep1	*Met Lys Leu STOP Val Gly STOP*
NEG_Pep2	*Met Lys Leu STOP Ala Ala STOP*
NEG_Pep3	*Met Lys Leu STOP Cys Pro STOP*
NEG_Pep4	*Met Lys Leu STOP Val Tyr STOP*
NEG_Pep5	*Met Lys Leu STOP Ser Val STOP*
NEG_Pep6	*Met Lys Leu STOP Lys Val STOP*

Figure 3.4. In-frame stop codons engineered in each of the NEG_Peps using specific forward primers. NEG_Pep1-6 are shown, where truncated peptides will be produced upon induction due to the presence of two in-frame stop codons. Specific primers used are listed in methods (chapter 2.1.7).

Cloning and expression of sequences with stop codons was performed as detailed in 2.2.4, which gave rise to six new parallel populations (NEG_Pep1_Stop-NEG_Pep6_Stop), each expressing truncated versions of the respective parental peptides (e.g. NEG_Pep1-6) in study. Growth measurements of these strains show somewhat mixed results, but more frequently a phenotypic rescue is observed. Two candidates, NEG_Pep2_Stop and NEG_Pep3_Stop show a complete phenotypic rescue from the growth defect post induction (Figure 3.5), with no prolonged growth seen in these contrary to the candidates expressing the respective full length peptides (shown in Figure 3.3).

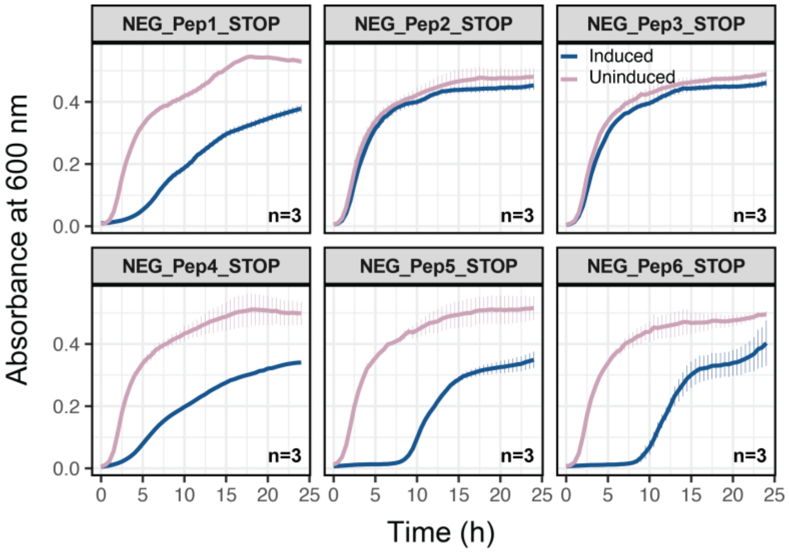

Figure 3.5. In-frame stop codons rescue the fitness defect fully or partially in NEG_Peps(1-4)_STOP but fail to show phenotypic rescue in NEG_Peps(5-6)_STOP. In-frame stop codons were introduced in the beginning of the random coding sequences assuming the formation of truncated peptides. Upon induction, the strains expressing truncated versions of the candidates show complete, partial or no phenotypic rescue (blue lines). Six stop codon candidates are shown in each box, where blue lines are the IPTG (1 mM) induced cultures

and pink lines are corresponding uninduced controls. All candidates are under the inducible vector pFLAG-CTC in *E. coli* K-12 DH10B strain background. Error bars represent standard error of the mean (SEM).

Consequently, it can be concluded that the fitness defect conferred by these candidates is likely due to a peptide-mediated mechanism, rather than the RNA. In the candidates NEG_Pep1_STOP and NEG_Pep4_STOP the truncated peptides only partially rescued the prolonged lag phenotype. On the other hand, NEG_Pep5_STOP and NEG_Pep6_STOP did failed to show a phenotypic rescue (Figure 3.5). These unsuccessful rescue effects could either be due to the expression of the residual short peptides, or a read through effect (failure in the recognition of the stop codon) or due to a mRNA mediated effect (where peptide production is not necessary to cause the fitness defect). Resolution of this question will require further experiments.

3.2.3. General stress responses activated in strains expressing NEG_Peps

To monitor the changes in the mRNA expression before and after induction of expression of NEG_Peps, a time series microarray experiment was performed (methods 2.2.12). Note that microarray was used due to a technical issue in performing RNA-seq experiments. The discontinuation of the bacterial ribosome depletion kit (monopoly of Illumina) led to an indefinite halt to RNA-seq experiments in bacteria until a new kit was standardized for optimal ribosomal RNA depletion. In a growing bacterial cell about 90% RNA is rRNA that needs to be removed for sensitivity to the remaining mRNA abundance. The total RNA extracted at three different time points was used for hybridization (see methods). Time point 1 consisted of non-induced cells in exponential phase, subsequent time point 2 consisted of cells induced for 1 hour and time point 3 consisted of cells induced for 16 hours. There were 5 replicates for each time point for two candidates (NEG_Pep1 and NEG_Pep5) and 3 replicates for empty plasmid control strains. Both of the NEG_Peps failed to show a full rescue in the peptide truncation experiment (using in-frame STOP codon, see Figure 3.5), implying that the host response might be to both, peptide as well as negatively acting RNA.

A principle component analysis (PCA) for expression data from the microarray analyses shows samples taken at three time points; T1, T2 and T3. Control (pFLAG) and test (NEG_Pep1 and 5) of T1 (non-induced) form a tight cluster with minimal variation in every principle component (PC). This was expected because the expression values in absence of NEG_Pep1 and 5 production (non-induced) should be similar to the pFLAG controls. PC1 explains the variability produced by different time points sampled (32.7% and 29.3% explained variance, Figure 3.6A. i. and B. i.). The control shows a clear distinction by time points along the PC1 axis (grey shapes). The clear spread in time points 2 and 3 is an indication of different gene sets being expressed in growing and stationary phase. PC2 and PC3 in both NEG_Pep 1 and 5 explain 12.6%, 21.3% and 11.4%, 8.7% respectively variance in the samples from the dataset, however each one explains variance of different features of the expression dataset. PC2 explains the variance observed one hour post induction in T2, in which the pFLAG controls (grey triangles) cluster away from the NEG_Pep (maroon triangles) producing samples (Figure 3.6A. i. and B. i., T2- induced 1 hour). T3 resolves better on PC1 as well as PC3 axis, separating the NEG_Pep5 producing samples from the pFLAG control and all other time points (Figure 3.6A. ii & B. ii., T3- induced 16 hours). Stationary phase related genes are expected to upregulate in all samples equally. It is observed for all NEG_Peps that the growth rates increase after 8-10 hours of lag and cultures attain maximum carrying capacity by 16-18 hours (see Figure 3.2A). The cultures presumably cope with the deleterious molecules by this time using an unknown mechanism.

Analysis of time series microarray data of the strains expressing NEG_Pep1 showed enhanced expression of genes associated with activation of the general stress response (raw data provided as supplementary files). Several operons known to be involved in peptide degradation and transport to the outside of the cell showed increased expression compared to pFLAG control in addition to the non-induced controls. The expression profiles of top 100 genes (sorted by highest fold change [log2FC] values) showed no overlapping genes to the corresponding non-induced candidates. The pFLAG control also showed no overlap with induced

NEG_Pep in the respective top 100 high fold change genes.

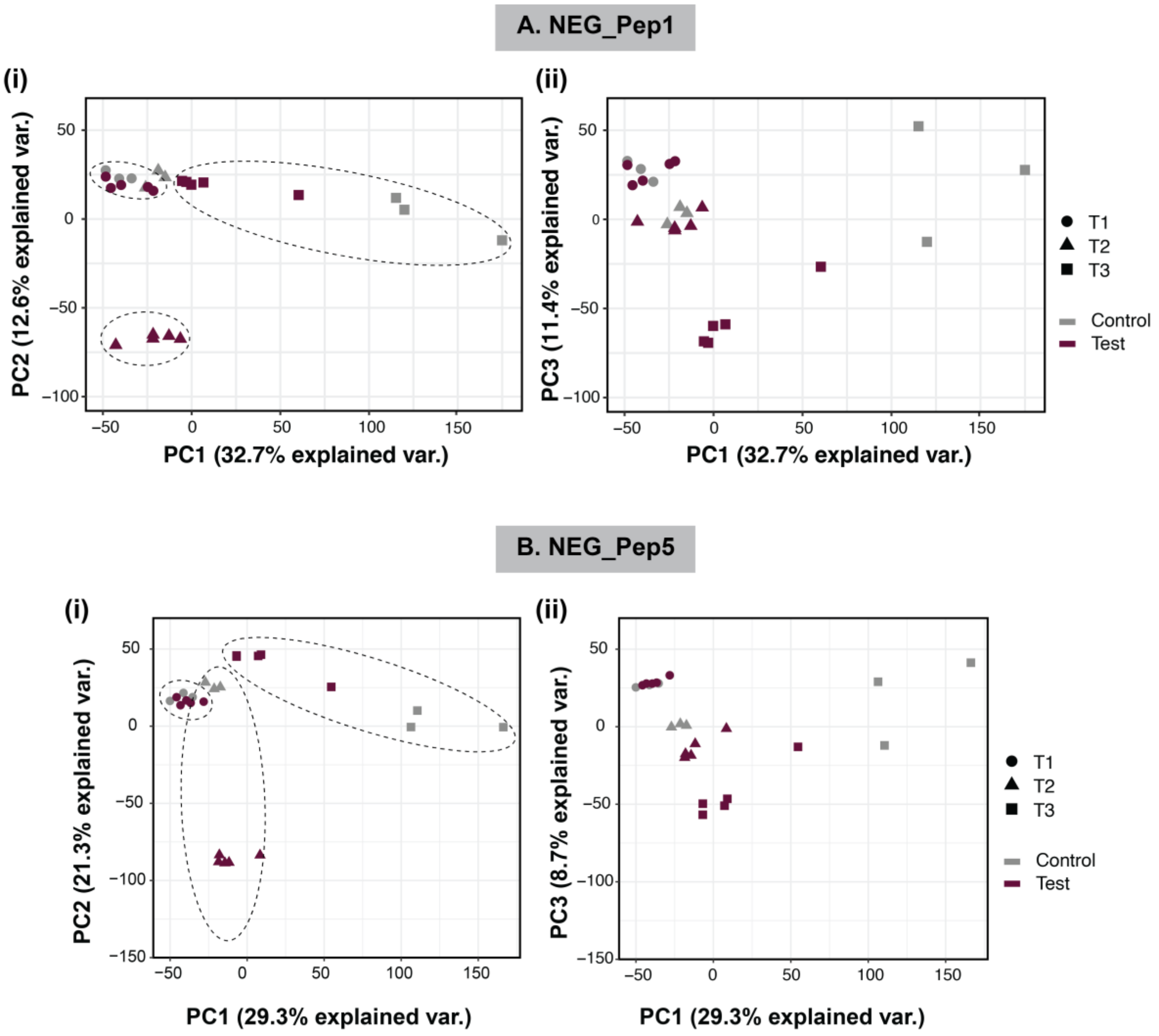

Figure 3.6. PCA of expression data of NEG_Pep1, NEG_Pep5 and pFLAG control strains at three time points (T1-3).

The principle component analysis (PCA) of all samples (15 test- NEG_Pep1 & 5 and 9 control- pFLAG) shows three clusters according to the different time points (ellipses- in greyscale) and within the three clusters the samples also resolve by NEG_Pep1 and control (maroon and grey points). At the non-induced time point 1 (circles), all samples (control and test) cluster together indicating similar expression values. A and B. i) PC1 versus PC2 plot shows that the variation within T2 (triangles; induced for 1 hour) is due the differences in the expression of test (maroon) and control (grey) explained by the PC2 axis. A and B. ii) PC1 versus PC3 plot shows clearly the samples at T3 (squares; induced for 16 hours) separating on the PC3 axis showing variation in their expression values.

In the specific case of NEG_Pep1 strains, at the second time-point, i.e. one hour after induction had undergone a massive change in their expression pattern (Figure 3.7A) when compared with pFLAG controls at the same time-point (Figure 3.7B). Several genes that were unique to NEG_Pep1 expressing strains are known to be associated with degradation of misfolded proteins and stress response. For example, the gene trio *dnaJ-grpE-dnaK* (average log2FC > 2.2) showed increased expression in NEG_Pep1 post induction (Liberek et al., 1991). DnaJ (a co-chaperone) along with GrpE (a heat shock protein) are known to activate ATPase activity of DnaK (a heat shock protein Hsp70, also a chaperon), which is known to be involved rescuing misfolded proteins and ensuring proper secretion outside the cells (Skowyra et al., 1990; Wild et al., 1992; Schroder et al., 1993; Ziemienowicz et al., 1993; Wild et al., 1996). A member of Hsp90 family of chaperons, HtpG (log2FC = 3.8) that is known to assist DnaK (Genest et al., 2011) showed enhanced expression in peptide producing strains. Another gene that showed increased expression was *htpX* (log2FC = 2.9). It encodes another heat shock protein and is known to assist with the degradation of misfolded proteins (Kornitzer et al., 1991; Sakoh et al., 2005).

In addition to *dnaJ-grpE-dnaK*, the genes of the phage shock operon (*pspABCDE*; average log2FC = 4.2) that are generally found to be expressed in the late stationary phase of cell growth (Model et al., 1997) were found to prematurely show over-expression in NEG_Pep1 post induction. All Psp proteins are involved in the stress response system of *E. coli* sp. (Huvet et al., 2011). Another gene *pspG* (log2FC = 2.9) known to regulate the Psp response was upregulated. Genes *pspA* and *pspG* although not physically in close proximity, have been shown to be involved in the maintenance of proton motive force (pmf) under stressful conditions by functioning as organised complexes (Engl et al., 2009).

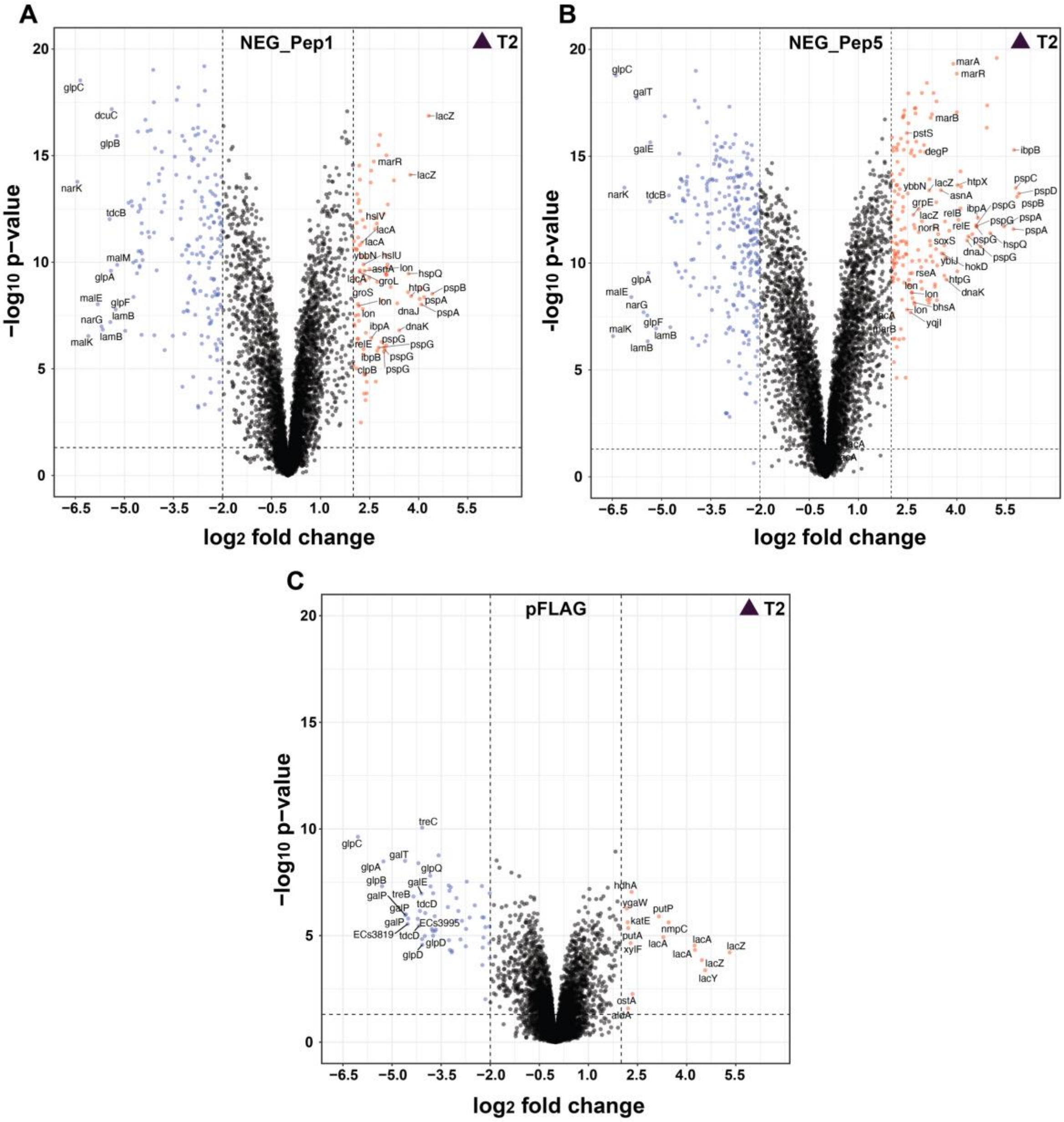

Figure 3.7. NEG_Pep1 and 5 show enhanced expression of various stress response genes after induction.

Volcano plots demonstrate the differential expression of genes in NEG_Pep1, NEG_Pep5 and pFLAG after 1 hour of induction (time point 2- T2) normalized with non-induced (T1) samples. T2 cultures were sampled in exponential phase; T1- obtained at OD_{600} of 0.4 and T2- after 1 hour of IPTG induction of the same culture. Genes above and below fold change of 2 and -2 (vertical dashed lines) are highlighted in red and blue respectively (p-values < 0.05, horizontal dashed line). A and B) Genes involved in stress response, efflux, etc. (labelled red data points) show enhanced expression in NEG_Pep1 and NEG_Pep2 respectively after 1 hour induction. C) pFLAG controls after induction show no significant differential expression of stress response genes. Genes of the *lac* operon (*lacA*, *lacY* and *lacZ*) serve as method control for IPTG induction. The red data points labelled with genes names are 20 stress related genes as categorised by gene ontology enrichment analysis. The blue data points labelled with gene names show the genes above log2 fold change of 4 (for representation). Expression data was analysed with R (version 3.6.3 and 4.0.0) using the Limma Bioconductor (v3.11) package (Smyth, 2005).

An elevated expression of the genes from operon *relBE* (average log2FC = 2.0) and the *lon* gene (log2FC = 2.3) were observed in both the induced candidate strains. RelBE is a toxin-antitoxin system where RelB acts as the antitoxin that is degraded by the Lon protease under starvation conditions (Christensen et al., 2001; Christensen and Gerdes, 2003). RelB degradations relieves repression of RelE (the toxin), and causes inhibition of translation. The gene *hokD* (host killing; log2FC = 1.8), a part of the transcriptional unit *relBE-hokD*, mediates plasmid stabilization by post-segregational killing (Gerdes et al., 1986). The polypeptide encoded by *hokD* was shown to kill host cells through a suicide mechanism, possibly by interfering with the host cell membrane as reported in one study (Gerdes et al., 1990). The transcripts of genes *lon*, *relBE* and *hokD* showed an enhanced expression in the induced candidate strains.

Additionally, the *marRAB* (multiple antibiotic resistance; average log2FC = 2.1) operon known to be involved in controlling several genes responsible for antibiotic resistance (Alekshun and Levy, 1997) and small multidrug efflux family genes (*mdtJI* operon; average log2FC = 3), showed elevated expression levels in induced candidates. A dual transcriptional activator, *soxS* (log2FC = 2.9), also known to be activated under stress and that has a function similar to the *marA* gene showed increased expression in induced candidates.

None of the genes described above showed enhanced expression in induced pFLAG control strains or the corresponding non-induced strains (Figure 3.7C, log2FC < 1.5). Majority of genes that showed enhanced expression in the induced NEG_Peps have been reported to have involvement in several biological processes like general stress response, multidrug efflux response and suicide response, confirming the deleterious nature of the NEG_Peps (Figure 3.7A). All the above mentioned genes also showed enhanced expression in NEG_Pep5 with similar or slightly higher fold change values (Figure 3.7B). Although in NEG_Peps1 and 5, only a partial phenotypic rescue was observed upon truncation of the peptides, it might be that the deleterious effect is caused by peptides rather than RNA. The stress response elevation is observed only after peptide induction suggesting the involvement of peptides rather than just RNA. It is likely that the over-expression leads to an immediate stress response and efflux activation allowing the cells to cope with the deleterious peptides instantaneously.

3.2.4. Suppressor screening and phenotype characterization

The frequency of suppressors was determined by directly plating fully grown NEG_Peps in LB+ Amp, on agar plates containing IPTG. The number of colonies that were visible (after 20-22 hours of incubation at 37°C) with over-expression of NEG_Peps was approximately five orders of magnitudes lower (about 10^4 CFU/mL) than the respective controls (about 10^9 CFU/mL) as shown in Figure 3.8. Usually an overnight culture grown on LB media has about 10^9 CFU/ mL after 20 hours. When peptide expression is not induced, the populations reach their optimal cell density (light pink box; Figure 3.8). Optimal population density is also reached in the pFLAG controls (black and grey plots, Figure 3.8). This shows that in absence of accumulating peptides, bacteria are able to divide and grow optimally but fail to do so in presence of NEG_Peps, leading to evolution of suppressors at a relatively high frequency (about 10^4 CFU/mL). The expression of NEG_Peps directly on solid media permitted isolation of suppressor-of-phenotype clones, which upon further investigation showed interesting characteristics (described in upcoming sections 3.2.4.1 and 3.2.5). A point to note here is that not all the colonies were suppressors-some colonies that grew at a normal size on IPTG-agar plates, failed to grow after the subsequent transfer. The prolonged lag phase may not just be a delay in growth of cells due to the over expression (Bolognesi and Lehner, 2018), but may as well be a peptide-mediated inhibitory effect that needs further investigation.

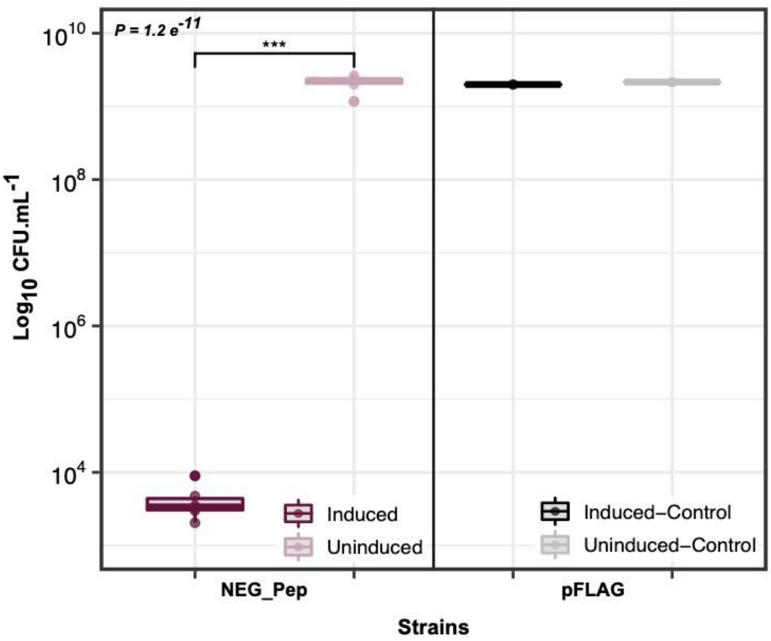

Figure 3.8. Apperance of suppressors on solid media containing IPTG. The colony forming units (CFU/mL) of induced NEG_Peps (1-6) show a significant decrease (t-test, p-value ***< 0.001) when plated on LB agar plates containing IPTG (for induction) compared to the corresponding non-induced (LB agar plates without IPTG) controls. The CFU/mL of the pFLAG control show a density of 10^9 CFU/mL. Maroon box plot shows $\log_{10}$ CFU/mL of six candidates (NEG_Pep1-6) on IPTG plates (LB+Amp+IPTG [1mM]) and the pink plot are the corresponding non-induced strains (LB+Amp). The adjacent panel shows pFLAG controls (in black) plated on induced plates together with non-induced plates (in grey).

Note that several punctate colonies come up after 22 hours but are not able to grow when subcultured on media containing ampicillin. These are most likely satellite colonies that come up after the antibiotic on the plate is degraded over time that may have lost the plasmid all together. I selected three full-sized colonies for further investigation from each of the NEG_Pep1-6 and tested their growth and expression in liquid media under induction. I hypothesized that the clones that arose to form visible colonies by 22 hours (on IPTG-agar plates) might have gained a mutation allowing the clones to adapt to the deleterious parental NEG_Peps1-6. In the following subsections, suppressor clones and their characteristics are described in detail.

3.2.4.1. Growth delay rescued in the suppressor clones

Suppressor colonies that exhibited phenotypic rescue were isolated from solid IPTG-agar plates (LA+ Amp+ IPTG). The cultures initially acclimatized to LB media with antibiotic, were plated on the IPTG-agar plates. The expression of peptides was enforced directly on the solid agar plates (with IPTG), which caused the emergence of evolved genotypes. Growth measurements of these evolved types showed a rescued phenotype, i.e. a phenotype without a prolonged lag phase similar to that of pFLAG controls. All suppressor clones (Figure 3.9) show phenotypic rescue i.e. absence of the prolonged lag phase as in their respective parental types NEG_Peps1-6 (see Figure 3.2). This phenotypic rescue was observed to be heritable (i.e. it was not a switcher phenotype) as subsequent growth experiments with the clones did not show any fitness disadvantage. The suppressor clones had successfully adapted to the peptide over-production without losing the plasmid or insert (this was tested using Sanger sequencing after plasmid extraction).

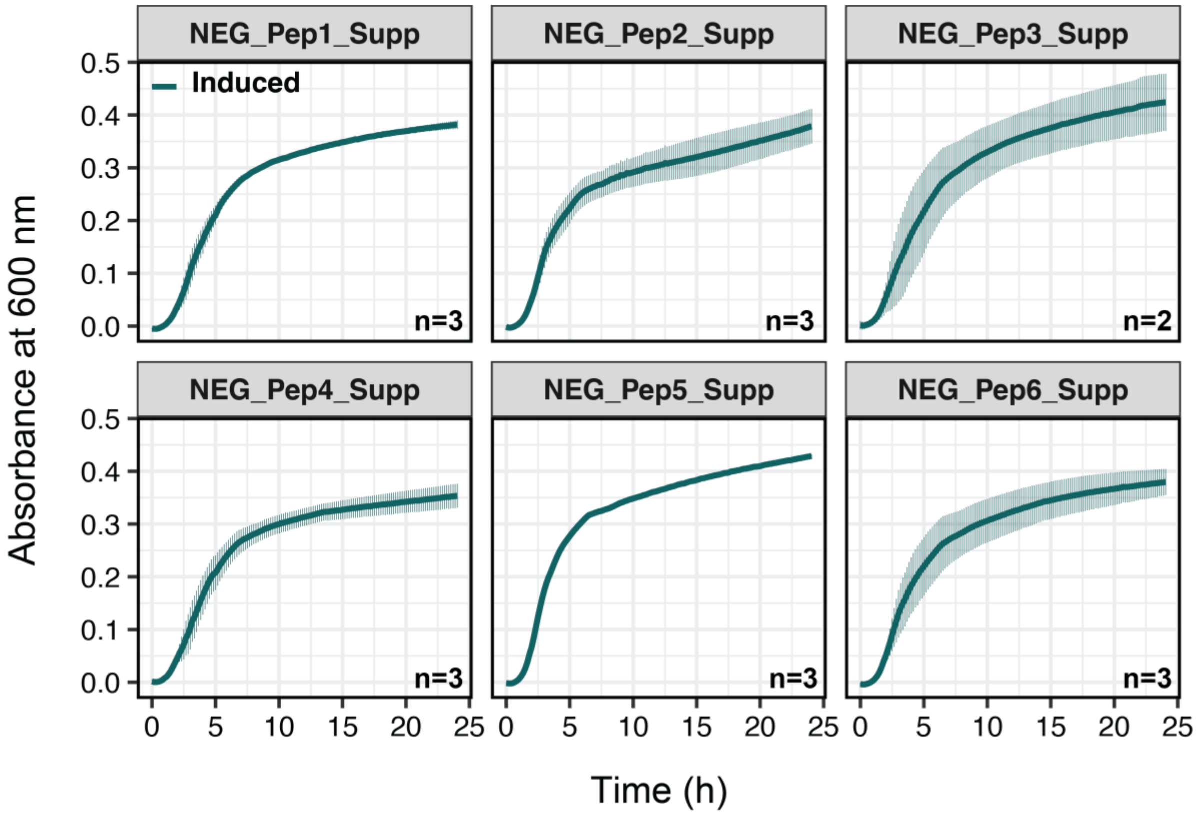

Figure 3.9. Suppressor clones selected from solid media that show a rescue of growth lag.

Green curves represent suppressor clones from corresponding parental strains NEG_Peps1-6 that show no growth defect. Growth measurements shown as OD versus time, performed in LB+ Amp+ IPTG at 37°C, 250 rpm for 24 hours. All samples shown are induced with 1 mM IPTG. Error bars are SEM. 'n' is the number of colonies from respective cultures of the same experiment.

The above observation implied presence of mutations in the genomic background of the suppressor clones. The suppressors appeared to have found a mechanism to reduce the negative effects from expression of the deleterious NEG_Peps. I hypothesized that a mutation, if present, would cause a decrease in the deleterious effects by one or more of the following mechanisms: (i) decrease in the expression of NEG_Peps, (ii) mutation in the interaction partner of NEG_Peps (iii) degradation of NEG_Peps, or (iv) increase in the efflux activity. Any of these alternatives would be expected to reduce deleterious effects of NEG_Peps inside cells, thereby making them tolerant or entirely resistant to them. To distinguish between these possibilities, the first step was to check the expression of peptides in the suppressor types and compare them with their corresponding parental clones.

3.2.4.2. Variable NEG_Pep expression in different suppressor genotypes

Evaluation of the phenotype of suppressors showed that there was a rescue from the prolonged lag phase where isolated clones managed to overcome the deleterious effects of peptides. The clones that showed this rescue showed a heritable change in the phenotype. These clones were classified as suppressors only when they showed a heritable phenotypic rescue. In order to confirm the presence of the full-length peptide expression in suppressor clones, western blot was performed taking advantage of the C-terminal FLAG-tag present at the tail ends of all candidates. The easiest indicator of whether peptide production had decreased was by estimating presence/absence of the peptides. All suppressors showed peptide expression, albeit not in the same way (Figure 3.10). The differences in the expression were found to be a result of distinctive underlying mutations in the backgrounds of the specific suppressors (discussed later 3.2.5.1). Specifically, NEG_Pep3 suppressors show extremely low levels of peptide production post induction. In NEG_Pep4, induction produces several non-specific bands, but the expression of peptide is evident in induced samples. The reason for the non-specific bands needs further investigation.

The following routes of adaptation were hypothesized for the emergence of suppressor types:

a. Mutations in the promoter region preceding the insert in the plasmid in order to reduce peptide over-expression.
b. Mutations in the insert sequence itself, producing different peptides (or truncated peptides if non-sense mutation) which may no longer have the associated fitness cost.
c. Loss of the entire plasmid (as a result of incorporation of the antibiotic resistance gene into the bacterial chromosome).
d. Mutations in the plasmid backbone.
e. Mutations in the genomic background.

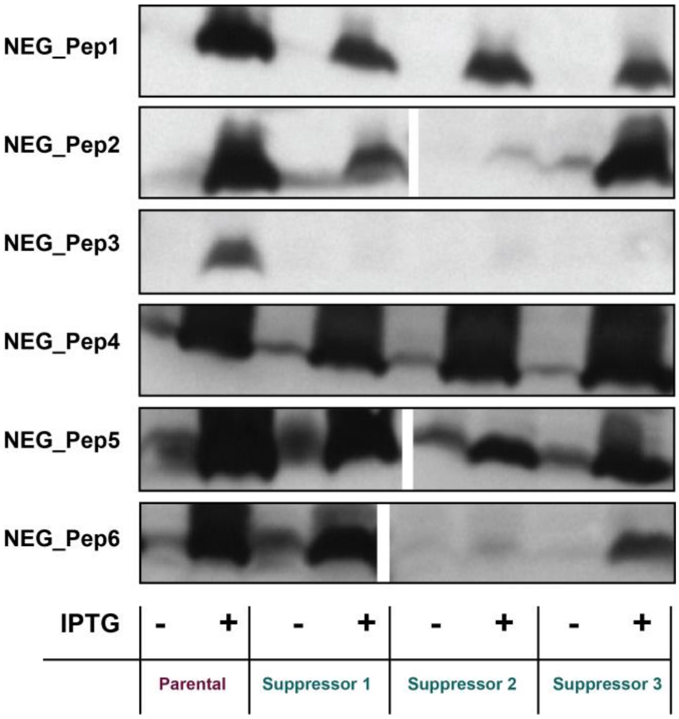

Figure 3.10. Peptide expression pattern changes in suppressors. All suppressor clones show presence (note that in NEG_Pep3 suppressors, faint bands are present) of peptides. The bands are approximately at 10 KDa which is the expected size of the 65 residue long peptides. The blot shows six NEG_Peps (in rows) and their corresponding suppressor clones (in columns) with and without IPTG induction. Note: there is faint expression in some of the uninduced samples which might be due to sample spill over from the adjacent wells. All inserts have a FLAG-tag at the C-terminal and are stained with mouse monoclonal anti-FLAG primary antibody, followed by a goat anti-mouse secondary polyclonal antibody with alkaline phosphatase for colorimetric detection (methods 2.2.10). This blot provides only a qualitative information and is not intended for quantitative interpretations.

By the method of elimination, I attempted to determine, which mechanism explained the evolution of suppressor-of-phenotype clones best. If the plasmid containing the inserts were completely lost from the host or if the insert sequence itself acquired a mutation, it would not be interesting in the light of questions asked in this study (since the NEG_Peps are on the plasmids). Hence, I proceeded to identify the presence of mutations in the promoter region or in the insert directly, which would shed light on the source of mutations. I performed Sanger sequencing using outer primers (see methods 2.1.7) with each of the suppressor clones to check whether the respective NEG_Pep sequences were preserved. Sanger sequencing revealed that all suppressor clones had the inserts intact and that none of the sequences had any mutation in them (data not shown here). The promoter regions of the insert sequences also did not harbour any mutations (Sanger sequencing data not shown).

3.2.5. Whole genome re-sequencing of the suppressors

The whole genome re-sequencing approach was taken to understand the evolution of suppressors of phenotype clones compared to the NEG_Pep parents. A total of 36 suppressor clones, three for each of the six parental candidates, in two different genetic backgrounds (DH10B and MG1655) were sequenced using the Illumina next generation sequencing platform (methods 2.2.13). Polymorphisms were identified using a manual pipeline (2.2.13.2) as well as an already available pipeline – breseq v0.35.1 (Deatherage and Barrick, 2014). A complete summary of mutations predicted can be found in Table 3.2.

3.2.5.1. Diverse mutations identified in suppressor clones

Custom reference genomes for each candidate strain backgrounds were created by using the re-sequenced NEG_Pep parental candidates (six, plus the pFLAG control), where all mutations in each of the parent were applied on to the reference file from NCBI before using them as references for assembling corresponding suppressor clones. This listed only the mutations that were unique to the suppressor clones, removing all common mutations in the parental strains. Whole genome re-sequencing provided hints about possible mutational solutions that suppressor clones had acquired to compensate the growth defect. Different NEG_Peps showed different strategies of adaptation on the genomic level (see summaries in Table 3.2 and Table 3.3).

Several mutations were found in the suppressor clones. When the same gene was seen to be targeted by a mutation in all three suppressor clones, it generally had mutated in the same position in all clones. This indicated that the mutation was already present in the initial population, albeit in low frequency and that it did not evolve in the suppressors independently. Two genes, namely *pcnB* (encodes a poly(A) polymerase) and *lacI* (encodes the lac repressor) were frequently hit with either insertion-deletion or substitution mutations (Table 3.2 and Table 3.3) are discussed exclusively in Chapter 1.

In DH10B background, the three suppressor clones that evolved from parental NEG_Pep1, had a mutation in *ydbA*, which codes for a putative outer membrane protein thought to be a member of the Autotransporter (AT) family (Zhai and Saier, 2002). All three suppressors from parental NEG_pep2 (DH10B background) had a single base pair deletion in the coding region of gene *tnaA* (Table 3.2), causing a frame shift and hence possibly a loss of function. The *tnaA* gene codes for an enzyme tryptophanase. TnaA mutants are known to have implications in the tolerance phenotype (Atsumi et al., 2010) and persister cell formation (Vega et al., 2012). Together with the mutation in the *tnaA* gene, one of the suppressor clone (Supp2) had a non-sense mutation in the *caiT* gene which encodes the L-carnitine: γ-butyrobetaine antiporter. When the same NEG_Pep2 was tested in MG1655 background, the gene *yhjE* showed a deletion in two of three suppressors (in same position, see Table 3.3). In NEG_Pep3 of MG1655 background, all three suppressors had a non-synonymous substitution in CybB (cytochrome b561) protein. There was a 359 bp deletion in a prophage origin (CP4-6 prophage) gene, *yagF* (encodes for a D-xylonate dehydratase) in Supp2 of NEG_Pep4 of the DH10B background. A DNA binding transcriptional dual regulator, *dsdC* was interrupted by a 9 bp IS10 (insertion sequence) fragment in the Supp3 clone of the parent strain NEG_Pep4. In the three suppressors of NEG_Pep5, the intergenic region upstream of a ribosomal rRNA gene rrsH had a substitution. Two non-synonymous mutations were found in one clone (Supp1) of NEG_Pep5 in the genes *ompN* and *anmK*. OmpN is an outer membrane porin N and AnmK is a anhydo-N-acetylmuramic acid kinase. AnmK recycles 1,6-anhydro-N-acetylmuramic acid (anhMurNAc) from the murein and returning it to the cellular biosynthetic pathways (Uehara et al., 2005). AnmK mutants have shown no accumulation of anhMurNAc inside the cells but a rapid efflux has been shown to take care of this such that it gets released in the spent media. In Supp3 clone, which evolved from the NEG_Pep6 parent, *mntH* gene that encodes a symporter protein was found to have a non-synonymous mutation in the coding region.

At least three candidates (NEG_Pep1, 2 and 5) in DH10B background and two candidates (NEG_Pep2 and 3) in MG1655 background, showed the same mutational hits in the three suppressor clones, which hints towards the possible presence of cellular targets of the respective NEG_Peps. This needs further confirmation by reinserting the mutation back into the suppressor genome and testing the deleterious effects. Due to the deleterious effects of the NEG_Pep interaction, the evolved suppressors might have targeted the genes underlying the affected interacting partner in order to relieve the deleterious effect. Different solutions are seen to be explored by different NEG_Peps in order to evolve into the suppressor genotype. Although reduction in the plasmid copy number or regulating the peptide expression on plasmid seems to be the most frequent solution (see Chapter 1, other mutational paths are also taken to reduce the deleterious effects of the peptides. NEG_Peps presumably create a massive stress on the cellular fitness, forcing them to adapt and evolve strategies to relieve this burden.

Candidate	**Suppressors: Mutation list**			**Position**	**Mutation***	**Type†**	**Annotation**
	Supp1	**Supp2**	**Supp3**				
NEG_Pep1	*pcnB*	*pcnB*	-	133188	IS10 (+) +9 bp	coding (35-43/1398 nt)	Poly(A) polymerase

	ydbA	*ydbA*	*ydbA*	1558099	A→C	(pseudo)gene (1555/2518 nt)	Putative outer membrane protein
NEG_Pep2	*pcnB*	-	-	132597	IS10 (–) +9 bp	coding (626-634/1398 nt)	Poly(A) polymerase
	tnaA	*tnaA*	*tnaA*	3975876	Δ1 bp	coding (60/1416 nt)	Tryptophanase/L-cysteine desulfhydrase
	-	*caiT*	-	41922	C→A	E4* (GAA→TAA)	L-carnitine: γ-butyrobetaine antiporter
	-	-	*hybO/ yghW*	3,230,737	IS1 (+) +9 bp	intergenic (-11/+170)	Intergenic; hydrogenase 2
NEG_Pep3	-	*pcnB*	-	132856	Δ1 bp	coding (375/1398 nt)	Poly(A) polymerase
	-	-	*lacI*	1388492	G→A	H74Y (CAC→TAC)	Lac repressor
	lacI/trp (?)	-	-	1387475	+22 bp	intergenic (-3002/ +66)	Insertion; intergenic upstream tryptophan
NEG_Pep4	*pcnB*	-	-	132600	(GC)3→2	coding (630-631/1398 nt)	Poly(A) polymerase
			pcnB	132206	IS150 (–) +3 bp	coding (1023-1025/1398 nt)	Poly(A) polymerase
	-	-	*dsdC*	2553355	IS10 (–) +9 bp	coding (78-86/936 nt)	DNA binding transcription dual regulator
	-	*yagF*	-	258398-258756	Δ359 bp		CP4-6 prophage; D-xylonate dehydratase
NEG_Pep5	*gmhB/ rrsH*	*gmhB/ rrsH*	*gmhB/ rrsH*	199054	G→A	intergenic (+206/-157)	Intergenic; upstream ribosomal RNA
	-	*lacI*	*lacI*	1387886	T→C	T276A (ACC→GCC)	Lac repressor
	ompN	-	-	1527469	C→T	R193H (CGC→CAC)	Outer membrane porin N
	anmK	-	-	1810158	T→G	N192H (AAC→CAC)	Anhydro-N-acetylmuramic acid kinase
NEG_Pep6	*lacI*	*NM*	-	1387891	T→C	D274G (GAC→GGC)	Lac repressor
	-	*NM*	*pcnB*	132787	+G	coding (444/1398 nt)	Poly(A) polymerase
	-	*NM*	*mntH*	2587955	A→G	I188T (ATC→ACC)	Mn2+/Fe2+ symporter

Table 3.2. Mutations found in suppressor genotypes in DH10B background. All mutations identified using Breseq (versions v0.33.2 and v0.35.1) pipeline and confirmed using Geneious Prime. NS = non-synonymous substitution and NM = no mutations found. Conflicting coverage denoted by '?'. All insertion sequences (IS) have caused target site duplication due to the insertion process (hence duplications of few nucleotides are observed flanking the IS element), which is represented by '+' followed by the number of bases duplicated (e.g. +9 bp; IS10). *= exact changes w.r.t. the positions. † = type of the mutations.

Parent	**Suppressor genotypes**			**Position**	**Mutation***	**Type†**	**Annotation**
	Supp1	**Supp2**	**Supp3**				
NEG_Pep1	*lacI*	*NM*	*NM*	365,909	T→G	T276P (ACC→CCC)	Lactose operon repressor
	-	*NM*	*NM*				
NEG_Pep2	*NM*	*pcnB*	-	158,929	$(GCC)_{3\to2}$	coding (196-198/1398 nt)	poly(A) polymerase
	NM	*yhjE*	*yhjE*	3,673,446	Δ6 bp	coding (214-219/1323 nt)	MFS superfamily sugar transport 1 family protein
NEG_Pep3	*cybB*	*cybB*	*cybB*	1,490,450	A→G	S109G (AGC→GGC)	cytochrome b561
NEG_Pep4	*lacI*	*NM*	-	365,848	A→C	L296R (CTG→CGG)	Lactose operon repressor
	-	*NM*	*pcnB* ← / ← *gluQ*	159,154	IS*2* (+) +5 bp	intergenic (-28/+28)	poly(A) polymerase/glutamyl-Q tRNA(Asp) synthetase
	-	*NM*	*pbpC*	2,644,818	T→A	G318G (GGA→GGT)	penicillin-insensitive murein repair transglycosylase, inactive transpeptidase domain
NEG_Pep5	*NM*	*NM*	*NM*				
NEG_Pep6	*lacI*	-	-	365,914	T→C	D274G (GAC→GGC)	Lactose operon repressor
	-	*lacZ* ← / ← *lacI*	*lacZ* ← / ← *lacI*	365,633	T→C	intergenic (-104/+19)	beta-D-galactosidase/lactose-inducible lac operon transcriptional repressor
	-	*dnaT*	-	4,599,842	A→C	F42C (TTT→TGT)	DNA biosynthesis protein (primosomal protein I)

Table 3.3. Mutations found in suppressor genotypes in MG1655 background.
All mutations identified using Breseq (versions v0.33.2 and v0.35.1) pipeline and confirmed using Geneious Prime. *= exact changes *w.r.t* the positions. † = type of the mutations. IS = insertion sequence, IS insertion has a target site duplication, here 5 bp, on either side of the IS element. NM = no mutations could be found.

3.2.5.1. No mutations could be identified in eight out of 36 suppressor genotypes

No mutations could be found in one of the 18 re-sequenced suppressor clones in DH10B background, which was surprising since the phenotype is heritable (see Table 3.2 NEG_Pep6- Supp2). It cannot be ignored that re-sequencing using the Illumina platform, unlike several other long-read sequencing methods, falls short in identifying large structural variations or repetitive region identifications. This may be one of the reasons for inability in the identification of significant mutations in the two strains. Whole genome re-sequencing was also performed for the evolved suppressors in MG1655 background (six parent NEG_Peps plus 18 evolved). There was a higher number of re-sequenced populations (seven out of 18 suppressor clones) where in no mutations could be identified (Table 3.3). The evolved suppressors that had mutations were either in the genes coding for transporters or genes involved in plasmid regulation and control. Further investigation is needed for suppressor genotypes, where no mutations could be identified. Long read sequencing would reveal the presence of large structural variations that might not have been captured by the short read sequencing approach.

3.3. Discussion

The findings of this chapter confirmed that candidate random peptides have deleterious effects on *E. coli* upon expression. NEG_Peps showed fitness defects through a prolonged growth lag upon expression in three different backgrounds of *E. coli*. The deleterious effects of NEG_Peps are pervasive across different *E. coli* backgrounds. It was found that the truncated versions of the NEG_Peps were fully or partially able to relieve the prolonged growth defect in NEG_Peps1-4 (see Figure

3.5). The production of truncated peptides (premature stop codons after 3th and 6th amino acid) prevents the accumulation of their full-length equivalents, which are presumably responsible for the fitness disadvantage. Two candidates (NEG_Pep5 and 6) did not show a phenotypic rescue upon truncation, which suggested a possible RNA mediated effect. Further confirmation of RNA versus peptide effects of NEG_Peps remains to be done.

Gene expression analysis of two candidate strains (NEG_Pep1 and 5) induced with IPTG had significantly elevated expression of genes involved in stress response whereas the corresponding non-induced as well as the empty vector controls did not elicit such a response. The fact that both NEG_Pep1 and 5 showed a partial and no growth rescue (respectively) upon blocking the peptide expression, hints towards a possible RNA effect in addition to the effect of peptide over-expression. Although the stress response genes elevate expression only after peptide expression, further confirmation is essential to determine possible RNA involvement. Specifically, genes of the *psp* operon are known to be involved in persister formation were seen to be generated in these induced strains (Vega et al., 2012). The Psp system is known to respond to a variety of environmental stressors including the damage repair of inner membrane of the cell (Kobayashi et al., 2007). Apart from the stress response, few other genes, *marRAB* (multiple antibiotic resistance), *soxS* and *mdtJI*, had elevated mRNA levels in the induced strains. The transcriptional activators MarA and SoxS initiate expression of *marRB* and *soxR*, which are known for broad spectrum antibiotic resistance and superoxide resistance respectively (Martin et al., 1996). MdtJI belong to the small multidrug resistance (SMR) exporters that are mainly known to be involved in the excretion of accumulated polyamine- specifically spermidine (Higashi et al., 2008). The MdtJI belong to the super family of multidrug resistance transporters or MDTs that are known to confer resistance to broad range antimicrobials in *E. coli* (Saier, 2000). The overall response suggested a strong deleterious effect caused by the NEG_Peps on the physiology of the host.

The deleterious NEG_Peps create a strong selection pressure on the strains expressing them. This leads to the emergence of suppressor genotypes, which no longer have the fitness defect. The initial hypothesis for the deleterious effects of specific NEG_Peps was that they might be interacting with components of the host machinery. This interaction might hinder some vital cellular processes giving rise to the growth defect. Screening and re-sequencing suppressor genotypes (that evolve by compensating the growth defect) would reveal these targets. The evolved genotypes showed acquisition of different mutations in their genomic background. Suppressors genotypes from two candidates: NEG_Pep1 and NEG_Pep2 had acquired mutations in *ydbA* and *tnaA* genes respectively (Table 3.2). The mutations were present in all three suppressor genotypes isolated from NEG_Pep1 and 2 each. Whether the specific NEG_Peps interact with the proteins encoded by these genes needs further investigation. Interestingly, TnaA mutants have been reported to increase persister cell formation and control multicopy plasmids (Chant and Summers, 2007; Vega et al., 2012). The suppressors adapt to the deleterious NEG_Peps by mutations in various target genes. A frequency dependent effect could explain the occurrence of background mutations, as the bacteria rarely acquire mutations in the random sequence itself. The peptides are always produced with the intact primary sequence (confirmed using western blots as well as insert sequencing).

Random peptides have the potential to cause fitness defects possibly by targeting cellular components of the host. The over-expression of NEG_Peps causes activation of stress response which helps the host cope with the deleterious effects of NEG_Peps. Suppressor colonies isolated from induced agar plates show diverse modes of adaptation to the expression of different NEG_Peps. An interesting question remains, whether the purified candidate NEG_Peps can have direct (*in vitro*) inhibitory effects on the growth of different bacteria.

Chapter 4. Results II

Evolution of suppressors by regulation of random peptide expression

4.1. Introduction

In the previous chapter, it was shown that expression of NEG_Peps had systematic deleterious effects on the fitness of *E. coli*. The *E. coli* populations expressing NEG_Peps respond to the peptide induced stress by regulating expression as well as by evolving into new suppressor genotypes. Some of the suppressor genotypes showed mutations in genes, the products of which, are presumably the targets of the corresponding NEG_Peps. Interestingly, I found a recurring class of mutations, in the genes *pcnB* and *lacI*, which will be discussed in detail in this chapter. Both the genes directly reduce the expression of NEG_Peps, either by plasmid copy number (*pcnB*) or reducing the transcription (*lacI*) in the cells.

The first gene, *pcnB*, encodes a poly(A) polymerase I (PAP I) which is responsible for addition of poly(A) tails to the 3' end of certain small RNAs (sRNA), which makes them targets for degradation (Haugel-Nielsen et al., 1996; Regnier and Hajnsdorf, 2009). Polyadenylation by PAP I at the 3' end of certain RNA molecules has been shown to promote degradation through different RNases (O'Hara et al., 1995; Mohanty and Kushner, 1999; Maes et al., 2016). The gene has also been previously reported to be involved in the copy number control of the plasmids of ColE1 origin of replication (Lopilato et al., 1986; Liu and Parkinson, 1989; March et al., 1989). It was shown that polyadenylation accelerates degradation of RNAI, an antisense repressor of replication of the ColE1 origin plasmids (Xu et al., 1993). RNAI represses the replication by complementary binding to an RNAII molecule, which is a primer for initiation of replication (Tomizawa and Som, 1984). RNAI and RNAII are complimentary RNA molecules, synthesized from opposite directions of each other and are involved in regulation the copy number of the plasmids (Tomizawa et al., 1981). Hence, the degradation of RNAI indirectly governs the replication of the plasmids through RNAII (which serves as a primer), which, otherwise would be unavailable, due to the presence of the stable antisense RNAI (Figure 4.1). In case of an inactive PAP I, a reduced plasmid copy number is expected and this mechanism could be exploited by the bacteria if there is accumulation of toxic products from the plasmids.

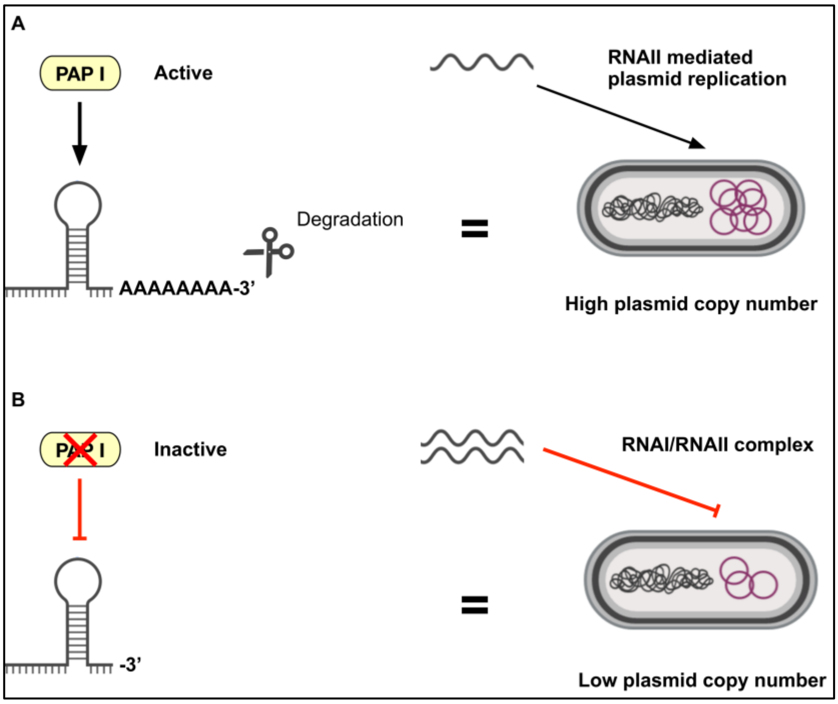

Figure 4.1. Mechanism of copy number control mediated by PAP I. A) When PAP I is active it polyadenylates the noncoding RNAI and allows the initiation of plasmid replication in cells through RNAII binding. B) When PAP I is inactive, the RNAI remains stable and binds complementarily to RNAII, thus reducing the initiation of plasmid replication leading to a low plasmid copy number in the cells. Graphic based on the review (Hajnsdorf and Kaberdin, 2018).

The second gene, *lacI*, encodes for the repressor protein of the lactose operon. LacI represses the lactose metabolism genes in the genome (in the absence of lactose). In the plasmids that are under *lacI* regulation, it represses the expression of gene of interest until the inducer molecule, IPTG (an allolactose analogue) is added. The repressor protein, LacI, is extensively studied *in vivo* and much of the effects of mutations are already known (Miller and Schmeissner, 1979; Gordon et al., 1988). Essentially, if the type of mutation and the position of mutation is known, one can readily predict the consequence it might have on the bacterial fitness.

Mutations that cause changes in the plasmid regulation (either copy number regulation or controlled expression of inserts) might be involved in reducing the overall concentrations of the deleterious peptides inside the cells. In order to understand the mechanisms underlying the targeted genes, it is necessary to identify the type and position of mutations on the gene. I hypothesized that the mutations in genes *pcnB* and *lacI* might be responsible for controlling the copy number and expression of peptides from the plasmid. Whole genome re-sequencing allowed me to identify suppressor genotypes that have a mutation in either of the two mentioned genes. With this information, I further investigated the type and position of each mutation in *pcnB* and *lacI*, to understand whether peptide mediated deleterious effects are concentration dependent. This would shed more light on the modes of action of these peptides inside the host cells.

4.2. Aims

1. To examine all evolved suppressors that have mutations specifically in *pcnB* or *lacI* genes in DH10B and MG1655 backgrounds using the WGS data.
2. To determine the position and type of mutations in the two mentioned genes.
3. To confirm the causative effects of the underlying mutations.
4. To test whether the deleterious effects of NEG_Peps are dosage dependent in *E. coli* (at a phenotypic level).

4.3. Findings

This chapter provides evidence of adaptation in suppressor genotypes through a reduction in the expression of NEG_Peps. The observations reported in this chapter not only shed light on the concentration dependant deleterious effects of the NEG_Peps (through mutations in LacI) but also on the role of a gene (*pcnB*) in plasmid copy number control.

4.3.1. Mutations in the *pcnB* gene affects plasmid copy number

Whole genome re-sequencing of three evolved suppressors clones from each of the six parental NEG_Peps in two different genetic backgrounds revealed, that six out of 18 suppressors in the DH10B background and two out 18 in the MG1655 background had mutations in the *pcnB* gene (Table 4.1). The *pcnB* gene encodes a poly(A) polymerase I (PAP I) that adds poly(A) tails to certain small RNA (sRNA) molecules targeting them for degradation (Mohanty and Kushner, 1999). All mutations found in *pcnB* were insertions and deletions (INDELs). Mutations were predominantly caused by insertion sequence (IS) elements interrupting *pcnB* together with a target site duplication of few nucleotides in each case. For example, suppressor 1 that evolved from NEG_Pep1 in the DH10B background had acquired an IS10 element in the coding region together with a duplication of nine bases flanking the insertion (Table 4.1). It is common to have some flanking bases duplicated due to the mechanism of the transposition, where the nicks are sometimes made on different sites on both the strands. Similarly, IS150 and IS2 were found to be interrupting the *pcnB* gene or directly upstream of the gene, presumably disrupting the promoter region, respectively in two independent suppressor genotypes. The IS elements that were found to interrupt *pcnB*, were IS150, IS10 and IS2, all of which have a size range of about 1.3-1.5 Kb and are quite large (Siguier et al., 2006). The suppressor genotypes with mutations in *pcnB* had presumably undergone a loss of function of the poly(A) polymerase (PAP I), since a majority of them were found to be frameshift mutations. The positions of each mutation identified within the reading frame is shown in Figure 4.2.

Background	**Parent**	**Suppressor genotype†**	**Postion in gene**	**Mutation***	**Type of mutation**
DH10B	NEG_Pep1	Supp1	35-43	Insertion IS10 +9 bp	Insertion + duplication
	NEG_Pep1	Supp2	35-43	Insertion IS10 +9 bp	Insertion + duplication
	NEG_Pep2	Supp1	626-634	Insertion IS10 +9 bp	Insertion + duplication
	NEG_Pep3	Supp2	375	Deletion 1 bp	Deletion
	NEG_Pep4	Supp1	630-631	Repeat GC deletion	Deletion
	NEG_Pep4	Supp3	1023-1025	GC deletion & IS150 +3 bp insertion	Insertion + duplication
	NEG_Pep6	Supp3	444	Insertion +1 bp	Insertion
MG1655	NEG_Pep2	Supp2	196-198	Repeat GGC deletion	Deletion
	NEG_Pep4	Supp3	Intergenic upstream (-28 bp)	Insertion IS2 +5 bp	Insertion + duplication

Table 4.1. List of mutations found in the *pcnB* gene of suppressors.

The mutations found in *pcnB* that encodes the Poly(A) polymerase (PAP I) are all insertions (with target site duplications) and deletions (INDELs). The insertions are a result of different IS element (IS10, IS150 or IS2) transposition and the deletions are mostly in the repeat region of the gene or in the intergenic region upstream of the gene. Mutation positions reported are the actual position on the gene and not w.r.t the entire genome. Supp1, 2 and 3 were genotypes isolated from growing the corresponding parent strain on plates with inducer IPTG. Candidates shown are either in K-12 DH10B or K-12 MG1655 backgrounds. † The genotypes that evolved from the original parent candidate listed (in the 'parent' column). * Mutation description in detail.

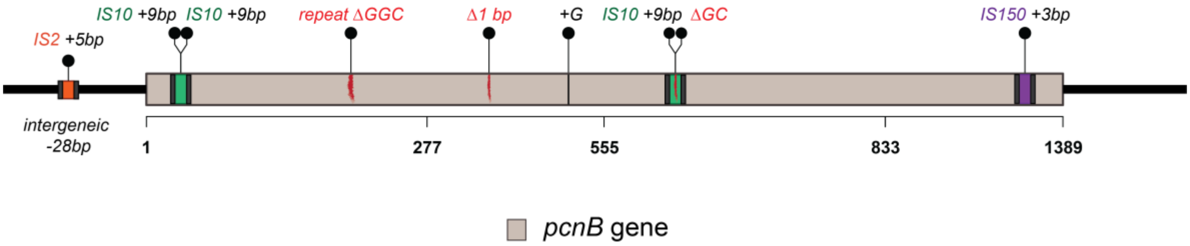

Figure 4.2. Positions of mutations found on *pcnB* gene in different suppressor genotypes.

PcnB gene block of 1389 bp is shown with all the suppressor genotype mutations marked on positions they occur in the gene from Table 4.1. Deletion is shown by a red mark, insertion elements IS2, IS10 and IS150 are shown in orange, green and purple respectively. The black blocks flanking the IS elements are the nucleotides that are duplicated in the target site. The black single line shows one base insertion. Upstream and downstream non coding segments are shown as black lines flanking the gene block.

4.3.2. Reduction in percentage of reads mapped to the plasmids of suppressors with *pcnB* mutation implies decreased copy number

Poly(A) polymerase I encoded by *pcnB* has been shown to regulate the plasmid copy number of the ColE1-like origin plasmids (Figure 4.1,(Lopilato et al., 1986; Liu and Parkinson, 1989)). To confirm if the mutations I found interrupting *pcnB* were reducing the plasmid copy numbers, I simply calculated the total reads that mapped to the plasmids in mutants and compared them with the parental plasmid reads containing strains. This served as a measure to assess the plasmid copy number by read depth. The read mapping was calculated by the number of reads that mapped exclusively to the plasmid, from the pool of total reads that map to both the genome and plasmid combined. This is a quick and easy approach to calculate relative percentage of reads that map to the plasmid and genome from a particular sequenced sample. The relative percentage of reads mapped to the each of the references was calculated using the following formula:

$$\%\, reads\ mapped\ to\ plasmid = \frac{Total\ no. of\ reads\ mapped\ to\ plasmid}{Total\ no. of\ reads\ mapped\ to\ plasmid + genome} * 100$$

The relative reads that mapped to the plasmids decreased to as low as 1% in the evolved suppressors that have the spontaneous INDEL in *pcnB* gene, as opposed to ~35-40% observed in the corresponding parental NEG_Pep1-6 candidates (Figure 4.3). The decrease in the percentage of reads that mapped to the plasmid reference was exclusively observed in those suppressor genotypes that had mutation in *pcnB*, not in the remaining suppressor genotypes (for e.g. see Figure 4.3, NEG_Pep1_Supp3 has 43% plasmid reads). Relative percentages of reads for NEG_Pep1-6 and three respective evolved suppressors (Supp1, 2 and 3) from each are shown in Figure 4.3. This gave the first confirmation that mutations in *pcnB* gene indeed influenced the plasmid copy number. In the MG1655 background, NEG_Pep6 suppressor genotype (Supp2), the relative reads aligned to the plasmid were only 0.5% as opposed to 17% in the others (see Figure 4.3B, NEG_Pep6_Supp2), but there was not underlying *pcnB* mutation involved (for mutation list refer chapter 3, Table 3.3). This specific suppressor had a mutation in *dnaT*, which encodes for DNA biosynthesis protein called primosomal protein I. It has been shown that DnaT is required for the replication of plasmids, especially for the pBR322 origin of replication (Masai and Arai, 1988, 1989). Hence, DnaT possibly plays the same role as PAP I to reduce the plasmid copy number. Next, I hypothesized that if the plasmid copy number decreases, susceptibly to the antibiotic marker must increase since the plasmids have an ampicillin marker. In order to test this hypothesis, I checked the growth of all the evolved suppressors on increasing concentrations of ampicillin containing agar plates using spot dilution assay (methods 2.2.11).

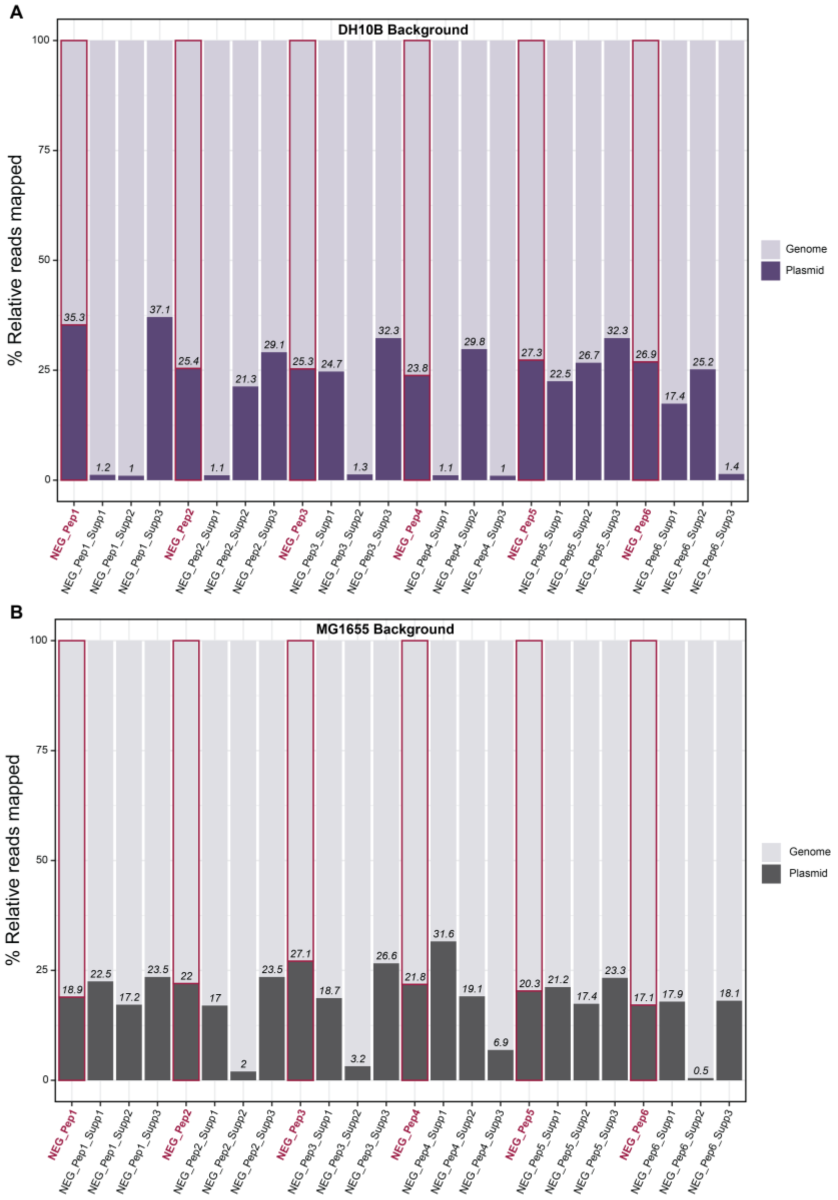

Figure 4.3. Percentage of reads mapped to the plasmid compared to its respective genome. Evolved suppressor genotypes that have mutations in *pcnB* show relatively less percentage of reads mapping to the plasmid reference compared others (compare with Table 4.1). Relative reads mapped to the reference plasmids and genomes of six parental candidates (NEG_Pep1-6, highlighted in maroon) with three suppressors for each parental strain are shown. **A)** DH10B and **B)** MG1655 backgrounds. Relative percentage of reads that map to the genomic reference (light bars) and plasmid reference (dark bars) were calculated from the total reads that map to both (hence relative). The percentages of plasmid reads are written on top of the darker bars. Number of reads mapping to the genome and plasmids were extracted from the read count summary after individual reference mapping using Geneious R11 (version 11.0.5).

4.3.3. Suppressors with mutation in the *pcnB* gene show increased susceptibility to ampicillin compared to their corresponding parents

Suppressors with mutation in the poly(A) polymerase revealed to have low number of plasmid reads, which indicated a low plasmid copy number. A low plasmid copy number would make these suppressor genotypes more susceptible to higher concentrations of ampicillin (Amp; all plasmids have an ampicillin marker). Genotypes with a high plasmid copy number (a characteristic of the pFLAG-CTC vector) should not show this trade off even at higher concentrations of ampicillin. A spot assay was performed with the previously sequenced six parental NEG_Pep (1-6) strains and three evolved suppressors (Supp 1, 2 and 3) corresponding to each of the parental strains.

The suppressor genotypes that had the mutated *pcnB*, indeed showed an increased susceptibility to ampicillin (Figure 4.4, marked with red asterisks). Suppressors clones from NEG_Pep1 and 2 showed an inhibition (on 250 µg/mL Amp concentration plates) at the 10^{-2} dilution, i.e., about 10^{5} cells in 10µL. But specific suppressor mutants in NEG_Pep3, NEG_Pep4 and NEG_Pep6, showed a 10 fold higher tolerance to ampicillin (250 µg/mL Amp) than others. They showed inhibition only at a dilution of 10^{-3}, i.e., about 10^{4} cells (in the 10 µL spot). Although this difference was observed on the antibiotic sensitivity assay, no marked differences were present in the plasmid copy numbers of these suppressors in the read mapping data (see Figure 4.3). Note that no inhibition was seen even at the highest antibiotic concentrations tested (250 µg/mL) in the pFLAG controls (see last panel of Figure 4.4) as well as the corresponding parental controls (first column of each plate in Figure 4.4). Since most mutations in *pcnB* are from the IS element insertion together with target site duplications, it is highly likely that these are loss of function mutations and PAP I gets inactivated in these mutants (see Figure 4.1).

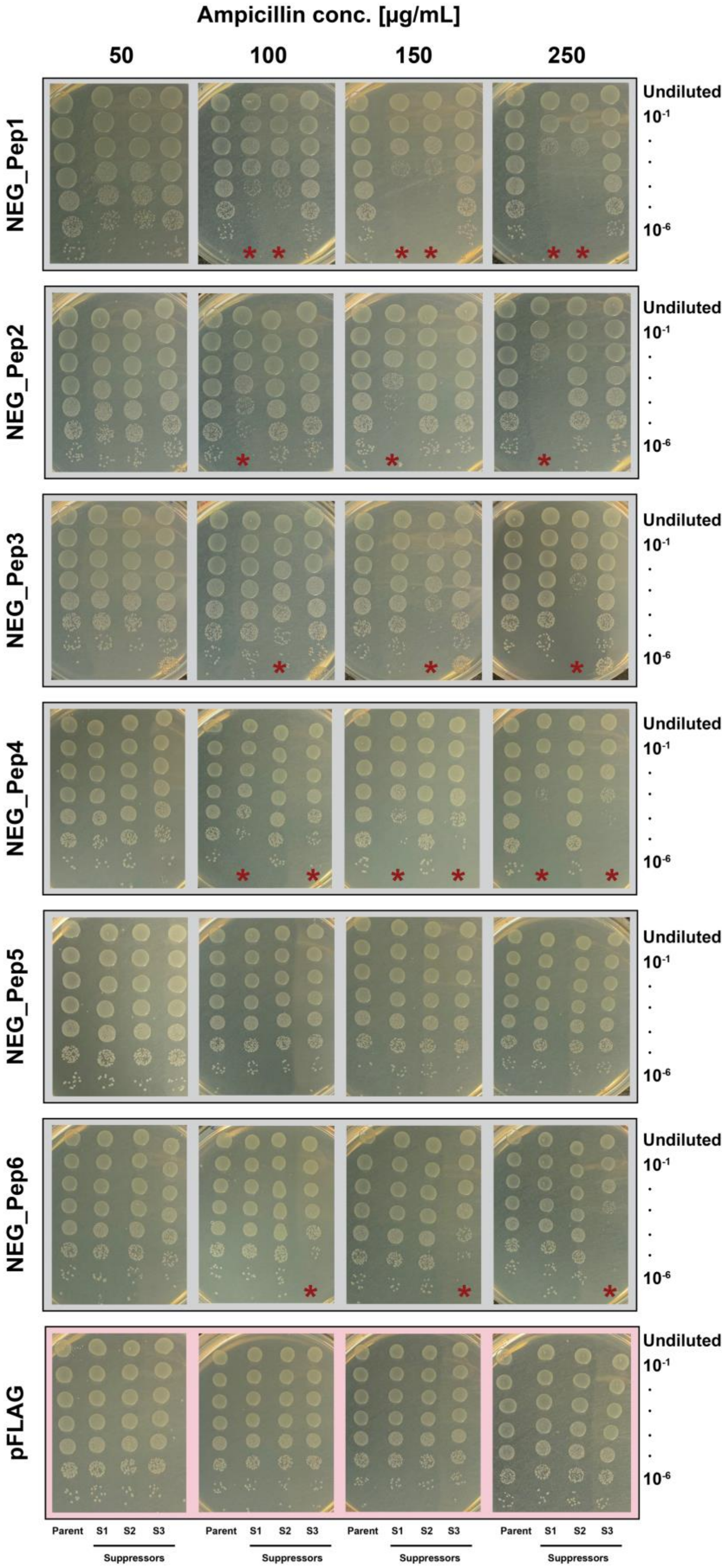

Figure 4.4. Mutation in *pcnB* gene reduces the plasmid copy number producing higher sensitivity to ampicillin. Suppressor genotypes that have a mutation in the *pcnB* gene (red asterisks) show increased sensitivity to ampicillin (>100 µg/mL). Control pFLAG parental strains as well as evolved genotypes do not show increase in sensitivity to ampicillin. Rows represent candidates (NEG_Pep1-6) plus control (pFLAG) and columns concentrations of ampicillin – low to high. Right hand axis shows increasing 10- fold dilutions of overnight cultures. Four columns inside every spotted box represent the parental strain, followed by three suppressor genotypes evolved from that parent (labelled in the bottom). Parent = parental NEG_Pep or pFLAG e.g. NEG_Pep1 and Supp1, 2 and 3 = corresponding suppressor genotypes. Red '*' represents suppressors

that have mutation in the *pcnB* gene. Cultures were spotted using 10µL of the appropriate dilution. Spot assay was performed on LB plates with four different ampicillin concentrations (from 50 to 250 µg/mL). Grey background is NEG_Peps1-6 and pink is pFLAG control.

Taken together, my data confirms the role of *pcnB* in the copy number control of pFLAG-CTC vector. Furthermore, in the evolved suppressors that have mutant PAP I, the reduction of plasmid copies directly curbs the absolute concentration of the deleterious NEG_Peps inside the cells. This in turn could cause the observed phenotypic rescue in these evolved suppressor genotypes. Mutants that manage to reduce the plasmid copy number arise as suppressors indicating that lower concentrations of NEG_Peps can possibly be tolerated. The deleterious effects of NEG_Peps might be present only at high concentrations such that it may be a concentration dependent response. Further experiments to validate the role of concentration of NEG_Peps are necessary to gain insights into the concentration dependent effects of NEG_Peps. I will discuss a preliminary experiment for this in the upcoming section **4.3.4**.

4.3.4. Candidates expressed in low copy plasmids do not confer fitness disadvantage to the host

Dosage sensitivity of NEG_Peps was tested by expressing the candidate NEG_Peps on a low copy (~ 20 copies/cell) plasmid vector, called pET45b (+) compared to the multicopy pFLAG-CTC (> 300 copies/cell). Three candidates (NEG_Pep1, 3 and 5) were ligated into the pET45b vector and transformed into the *E. coli* K-12 DH10B background. Phenotypes were investigated using growth curves performed as previously mentioned (2.2.7.1). The vector also had ampicillin as the antibiotic selection marker and an IPTG inducible promoter. There was no significant difference in the growth dynamics between induced and non-induced states (Figure 4.5A). In other words, the strains expressing the deleterious random peptides did not show the fitness defect (contrary to the high copy expression; see chapter 3, Figure 3.2) in the low copy vector. The growth rates (calculated as the maximum slope after calculating all slopes with sliding windows with 5 values using GrowthRates v4.2) show no difference in the induced versus the non-induced NEG_Pep1, 3 and 5 (Figure 4.5B). Although this is a preliminary evidence, it suggests a possible dosage dependent effect of the deleterious random peptides. Further experiments will be required for the actual quantification of the peptide concentrations inside the cells.

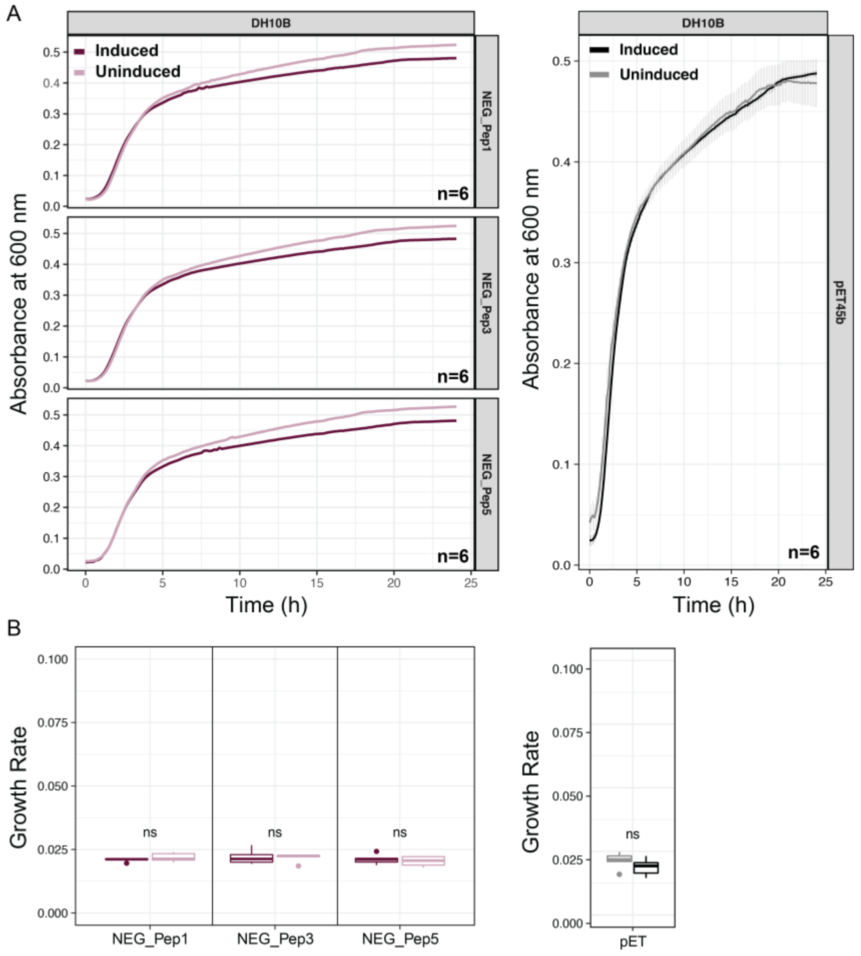

Figure 4.5. Candidate peptides expressing from the low copy (~20 copies/cell) vector do not confer growth disadvantage to the host. A) Growth curves of three candidates (NEG_Pep1, 3 and 5) engineered in low copy plasmid pET45b (+). Strains expressing the peptides have similar growth dynamics compared to strains that do not express the peptides and vector controls. Error bars represent standard error of the mean (SEM); lack of error bars is due to miniscule SEM values. ‘n’ is the number of replicates. B) Growth rates show no significant differences among induced and non-induced strains (t-test; p>0.05). Candidates induced with 1 mM IPTG are shown in maroon whereas the corresponding non-induced controls are in pink. Empty vector controls are represented in black when induced and grey when non-induced. Growth measurements were performed in LB + Amp medium at 37°C for 24 hours with OD reading taken every 10 minutes with 5 minutes shaking prior to OD reading at 600 nm.

4.3.5. Non-synonymous substitutions in the LacI repressor are found exclusively in the core of the protein

An important protein in the regulation of expression of the peptides is the repressor LacI. The NEG_Pep candidates in the vector pFLAG-CTC are under a strict repression through LacI, which is encoded on the vector. The repression gets relieved after addition of the inducer IPTG, which switches on the expression of the random insert downstream. Note that the genomic background also has *lacI* that regulates the Lac operon genes for lactose metabolism.

Genomic background	**Candidate parent**	**Evolved suppressor***	**Mutation position†**	**Reads aligned to new base/ ref base**	**Frequency**
DH10B	NEG_Pep3	Supp3	H74Y	43545/105	99.8
	NEG_Pep5	Supp2	T276A	46355/136	99.7
	NEG_Pep5	Supp3	T276A	42286/106	99.8
	NEG_Pep6	Supp1	D274G	22259/107	99.5
MG1655	NEG_Pep1	Supp1	T276P	24600/147	99.4
	NEG_Pep2	Supp3	N125Y	34758/204	99.4
	NEG_Pep4	Supp1	L296R	32315/125	99.6
	NEG_Pep6	Supp1	D274G	21359/115	99.5

Table 4.2. Mutations found in the *lacI* genes of the suppressor genotypes of two genetic backgrounds.

All mutations found were non-synonymous substitutions in the core region of the repressor protein LacI. All mutation in *lacI* gene that were found in the whole genome re-sequencing data of six candidates in two genetic backgrounds are listed. The number of reads with new base and reference base (ref) are shown together with the frequency of the observed mutation (calculated as (new base reads/total reads)*100). The candidates were on a pFLAG-CTC vector in either K-12 DH10B or K-12 MG1655 backgrounds. The genotypes that evolved from the respective candidate listed (in the parent column). † Mutation position w.r.t the LacI protein (360 amino acids) with the change in specific amino acid represented as flanking letters.

In each of the two genetic backgrounds DH10B and MG1655 tested, four out of 18 re-sequenced evolved suppressor genotypes had a mutation in the *lacI* (listed in (Table 4.2). The LacI protein is functional as a tetramer (Figure 4.6A for 3D structure) where each monomer consists of an N-terminal head piece (1-59 aa) and a C-terminal core (60-360 aa) (Lewis et al., 1996). The C-terminus has two important domains which are: the DNA binding domain and the inducer or sugar binding domain (Wilson et al., 2007) (Figure 4.6B). The sugar binding domain is a cleft where allolactose or IPTG binds such that the repressor gets released and the expression of downstream genes begins. All mutations in LacI repressor were found to be non-synonymous substitutions (Table 4.2). Upon examining the positions of these substitutions, it was found that all substitutions were in the core (Figure 4.6B) of the repressor protein (Markiewicz et al., 1994). It has been shown in previous studies that the core of the LacI repressor is involved in tetramer formation (Schmitz et al., 1976) as well as inducer binding (Miller and Schmeissner, 1979). Mutations in the amino acid residues that are involved in the sugar binding (I^s type mutants) have previously been analysed (Suckow et al., 1996). The specific substitutions I found (except for H74Y) have been shown to be involved in the inducer binding (Wilson et al., 2007) which, in my experiments would interfere with the IPTG binding to the LacI, subsequently keeping the repressor bound to the operator. The addition of IPTG can no longer induce the expression from the vector causing the reduced expression of NEG_Peps leading to the automatic weakening of deleterious peptide accumulation.

Although *lacI* is present on the genome and the plasmid, I could identify the source of the mutation through the read depth of my re-sequencing samples. All the LacI mutations showed more than 98% reads that mapped to the mutant nucleotide (alternate allele) whereas the remaining reads mapped to the reference (see Table 4.2 frequency column). The genome has only one copy of *lacI*, whereas the plasmid would have as many as their copy number inside a cell. Given that almost all reads have the mutation, it is extremely likely that the mutation arose in the plasmid copy of *lacI* rather than the genomic one. Multicopy plasmids provide a much larger mutational target size compared to the single copy present in the chromosome, which together further corroborate my findings.

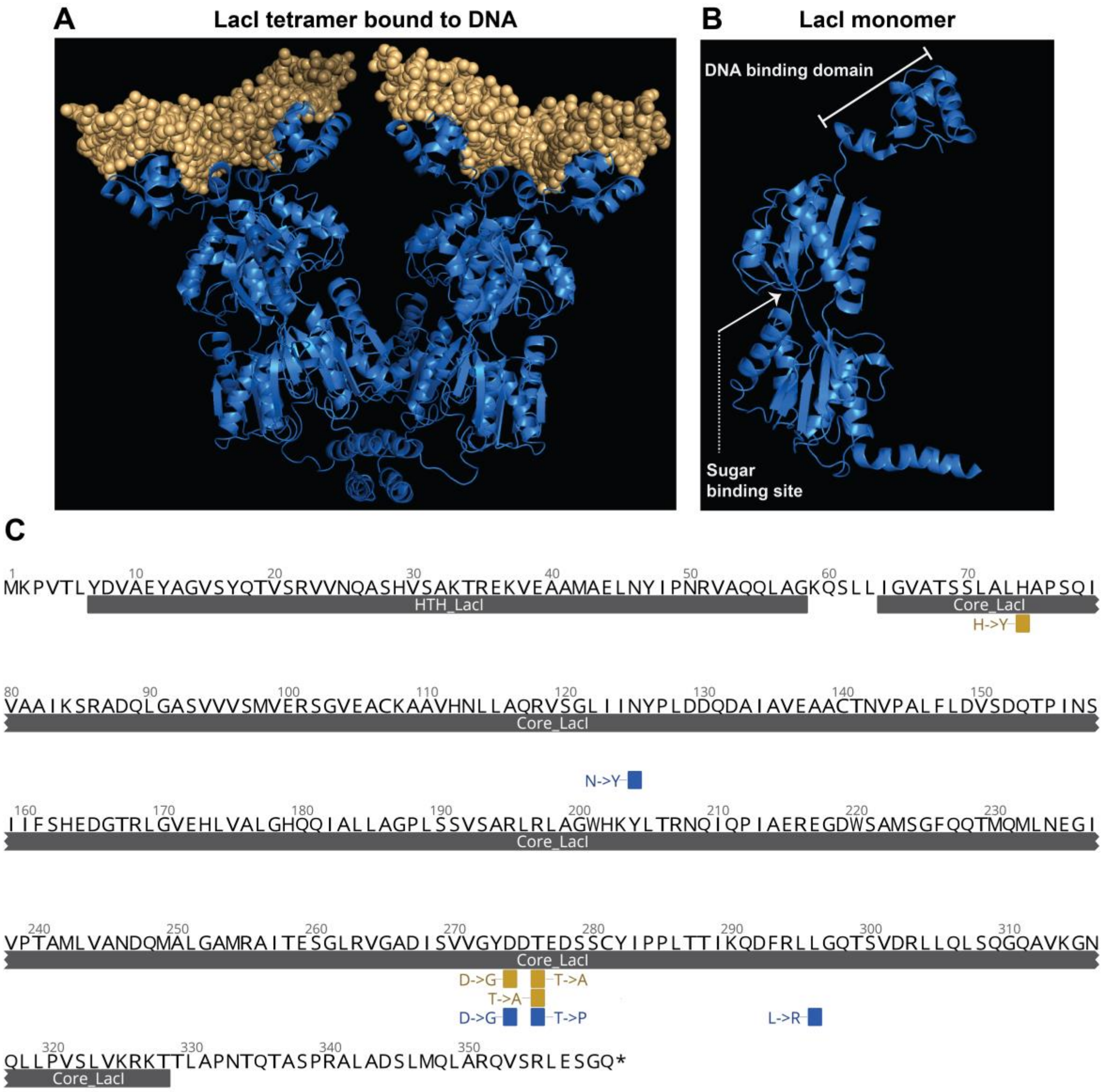
A
LacI tetramer bound to DNA
B
LacI monomer
DNA binding domain
Sugar
binding site
C
MKPVTLYDVAEYAGVSYQTVSRVVNQASHVSAKTREKVEAAMAELNYIPNRVAQQLAGKQSLLIGVATSSLALHAPSQI
1
10
20
30
40
50
60
70
HTH_LacI
Core_LacI
H->Y
VAAIKSRADQLGASVVVSMVERSGVEACKAAVHNLLAQRVSGLIINYPLDDQDAIAVEAACTNVPALFLDVSDQTPINS
80
90
100
110
120
130
140
150
Core_LacI
N->Y
IIFSHEDGTRLGVEHLVALGHQQIALLAGPLSSVSARLRLAGWHKYLTRNQIQPIAEREGDWSAMSGFQQTMQMLNEGI
160
170
180
190
200
210
220
230
Core_LacI
VPTAMLVANDQMALGAMRAITESGLRVGADISVVGYDDTEDSSCYIPPLTTIKQDFRLLGQTSVDRLLQLSQGQAVKGN
240
250
260
270
280
290
300
310
Core_LacI
D->G
T->A
T->A
D->G
T->P
L->R
QLLPVSLVKRKTTLAPNTQTASPRALADSLMQLARQVSRLESGQ*
320
330
340
350
Core_LacI

Figure 4.6. 3D structure and positions of substitutions in the Lac repressor.
A) 3D structure of the Lac repressor tetramer bound to the DNA bound is shown (generated in pyMOL v2.3.5 using PDB file '1LBG', modified from (Wilson et al., 2007)). **B)** Monomer of LacI with the DNA binding domain and the sugar binding cleft is shown. **C)** Amino acid sequence of LacI with the positions of substitutions found in suppressor genotypes of DH10B (in blue) and MG1655 (in yellow) backgrounds. All mutations are in the core region of the protein.

4.4. Discussion

The results in this chapter suggest that the mutations in *pcnB* and *lacI* are indeed responsible for regulating the effective concentration of candidate deleterious peptides in the described suppressor genotypes (see Figure 4.7 for summary). When three NEG_Peps (1, 3 and 5) were expressed from a low copy plasmid, the deleterious fitness effect no longer persisted (no growth lag). This indicates a possible concentration dependent mode of action, i.e. a higher concentration of the deleterious peptides (or RNA) are required to produce fitness defects.

Several IS elements that were found to interrupt *pcnB* (encodes poly(A) polymerase I or PAP I) most likely caused gene inactivation- the most commonly observed consequence of an IS transposition (Parkhill et al., 2003; Vandecraen et al., 2017). The inactive PAP I further failed to polyadenylate the sRNAs, leading to plasmid copy reduction (see Figure 4.1). The reduced plasmid copy number would consequently result in a reduced production of the deleterious peptides, one that is likely bearable by the cells. Plasmid copy number reduction had a trade-off, which made the PAP I mutants more susceptible to increasing concentrations of ampicillin. The mutants could not survive high concentrations of ampicillin, since the mechanism of regulating the plasmid copy numbers was compromised. Especially since the *pcnB* mutations directly affect the plasmid copy numbers, it is likely that the negative effects of peptides are dosage dependent. Tuning the dose of peptides might be a potential adaptive solution by the cells under a strong selective pressure from deleterious peptides. Furthermore, concentration dependence was revealed to play a role when three candidate NEG_Peps that were engineered under the regulation of a low copy plasmid (~20 copies/cell) showed no fitness defect upon expression (see Figure 4.5), unlike the prolonged growth defect observed when expressed on a high copy plasmid (pFLAG-CTC > 300 copies/cell, chapter 3, Figure 3.2). Although this was a preliminary result, it gave an insight into a probable concentration dependent or a threshold sensitive mechanism of action.

Some of the evolved suppressor clones had mutations in the *lacI* gene, which codes for the repressor of the lactose operon. A copy of *lacI* is also present on the vector to regulate the expression of random inserts downstream, such that the expression only begins upon addition of IPTG (sugar analogue inducer). All the mutations in *lacI*, were single nucleotide changes leading to non-synonymous substitutions, present exclusively in the core of the repressor (LacI) protein (Table 4.2). The core region of the LacI repressor has been shown to be involved in tetramer formation and inducer attachment (Markiewicz et al., 1994). Previous studies have shown that LacI mutants that no longer respond to the inducer molecules (termed as I^s type mutants) can arise from specific substitutions in the core region of the protein (Gordon et al., 1988; Wilson et al., 2007). The non-synonymous substitutions in the core region were found to be in the inducer binding cleft, which has been shown to produce an unresponsive repressor, meaning it cannot be released from the operator site. In other words, LacI becomes unresponsive to the inducer molecule, IPTG, which ultimately triggers a repressed state of the insert gene downstream on the plasmid. If IPTG fails to bind to the LacI repressor, induction of expression would be halted eventually affecting the peptide production. This signifies a direct control of expression of the random peptides that are under control of the LacI repressor. Although the chromosome also possess *lacI*, the mutation was likely to be on the plasmid *lacI*.

Interestingly, *pcnB* mutants mostly comprised of large insertions of the IS elements whereas *lacI* mutants were exclusively single base change making a non-synonymous change in the amino acid residue. This is due to the fact that the insertions lead to inactivation of the gene whereas substitutions at least in this study lead to inactivation of particular domains. PAP I was presumably inactivated completely due to the > 1.5 Kb long insertions, leading to a loss of function. PAP I is known to be dispensable in *E. coli* (Masters et al., 1993). In case of LacI, inactivating the sugar binding ability while retaining the repressor function was critical for the cells to keep the expression of deleterious NEG_Peps curbed.

Mutations in both genes *pcnB* and *lacI* seem to be involved in reducing the peptide production by controlling either the plasmid expression or the number of copies. Suppressor mutants that had mutations in any of the two genes were able to increase in fitness comparable to the wild type strains (see chapter 3). Both gene products PAP I and LacI are directly or indirectly responsible for plasmid maintenance and control, but both ultimately control the effective peptide concentration in the cells. This points towards a rather general strategy of adaptation, one unlikely to be specific to the peptide sequences. Overall, these findings highlight the probable dosage sensitive effects of random candidate peptides and diverse ways in which *E. coli* adapt to arrive at the same solution.

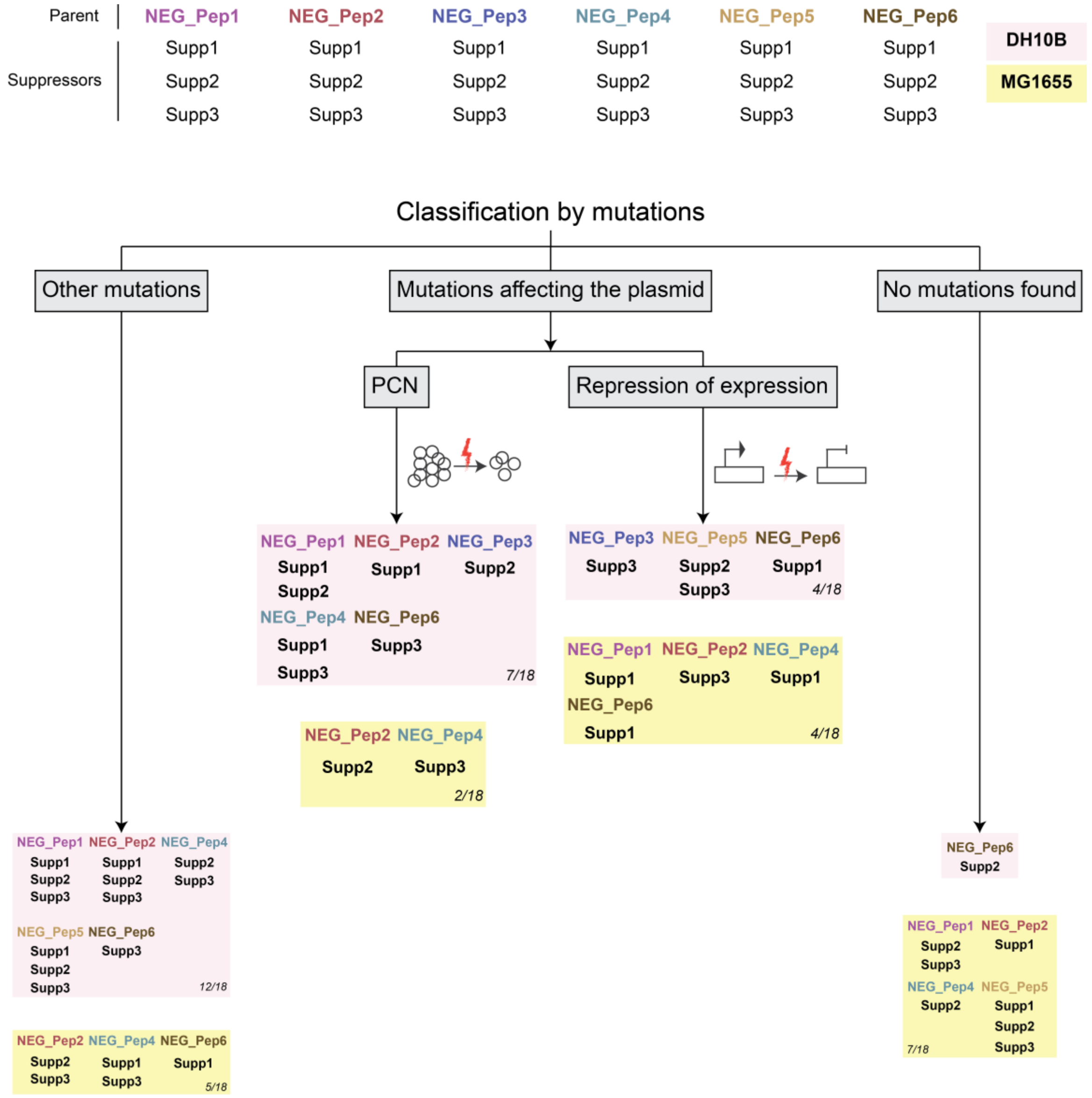
Parent
Suppressors
NEG_Pep1 Supp1 Supp2 Supp3
NEG_Pep2 Supp1 Supp2 Supp3
NEG_Pep3 Supp1 Supp2 Supp3
NEG_Pep4 Supp1 Supp2 Supp3
NEG_Pep5 Supp1 Supp2 Supp3
NEG_Pep6 Supp1 Supp2 Supp3
DH10B
MG1655
Classification by mutations
Other mutations
Mutations affecting the plasmid
No mutations found
PCN
Repression of expression
NEG_Pep1 Supp1 Supp2 NEG_Pep2 Supp1 NEG_Pep3 Supp2
NEG_Pep4 Supp1 Supp3 NEG_Pep6 Supp3
7/18
NEG_Pep2 Supp2 NEG_Pep4 Supp3
2/18
NEG_Pep3 Supp3 NEG_Pep5 Supp2 Supp3 NEG_Pep6 Supp1
4/18
NEG_Pep1 Supp1 NEG_Pep2 Supp3 NEG_Pep4 Supp1
NEG_Pep6 Supp1
4/18
NEG_Pep1 Supp1 Supp2 Supp3 NEG_Pep2 Supp1 Supp2 Supp3 NEG_Pep4 Supp2 Supp3
NEG_Pep5 Supp1 Supp2 Supp3 NEG_Pep6 Supp3
12/18
NEG_Pep2 Supp2 Supp3 NEG_Pep4 Supp1 Supp3 NEG_Pep6 Supp1
5/18
NEG_Pep6 Supp2
NEG_Pep1 Supp2 Supp3 NEG_Pep2 Supp1
NEG_Pep4 Supp2 NEG_Pep5 Supp1 Supp2 Supp3
7/18

Figure 4.7. A summary of all the candidates and their evolved genotypes classified into mutational types.

The summary focuses of the mutations that affect the plasmid replication or expression. The remaining strains are classified into other mutations (described in chapter 3) and no mutations where no identifiable mutations were found (using the Illumnia platform).

Chapter 5. Results III

Effects of expressing positive random coding sequences in Escherichia coli

5.1. Introduction

Organisms use diverse mechanisms in order to acquire novelty into their genomes (reviewed in (Kaessmann, 2010) and (Long et al., 2003)). *De novo* gene evolution is one such mechanism for evolution of new genes (Shaw et al., 1978). This is the only evolutionary mechanism, where innovability (the ability to create innovation) does not rely on pre-existing gene templates (as in the process of lateral gene transfer, gene duplication, gene fusion, etc.). Essentially, *de novo* gene emergence materializes after non-coding regions in the genome acquire the ability to transcribe and translate stretches of nucleotides. These spurious products are referred to as proto-genes, that are in the process of becoming a fully functional gene (Carvunis et al., 2012). This proto-gene stage has been mimicked in seminal studies, which have shown that random sequences with biological function can be selected *in vitro* (Ellington and Szostak, 1990; Famulok and Szostak, 1993; Wilson and Szostak, 1999; Keefe and Szostak, 2001; Seelig and Szostak, 2007; Felippes et al., 2008). More recently, it has been shown that random peptides can be selected to confer antibiotic resistance (Knopp et al., 2019). In Knoop, et al., the authors selected for random peptides that increased the aminoglycoside resistance of . Another study highlighted that mini-genes of 20-mer random peptides conferred increased resistance when grown under high stress compared to the controls that did not harbour the mini- peptides (Stepanov and Fox, 2007). Previous studies have shown that mutant strains (such as auxotrophs or strains lacking a vital protein) can be rescued by novel proteins screened from a random peptide library (Fisher et al., 2011; Donnelly et al., 2018). *De novo* peptides have been screened by the above approach in a previous study, where the authors applied strong antibiotic selection pressure and managed to screen three candidates that conferred resistance to several aminoglycoside antibiotics tested (Knopp et al., 2019). Random sequences and their impact on the fitness of host in absence of a direct selective pressure has not been studied.

Most studies to date have focused on a directed evolution (selective) approach to screen for novel biological functions of random sequences. In contrast, a previous study in our lab measured growth by competition between cells in a population expressing different random peptides, which also served as unique barcodes in the system for counting (Neme et al., 2017). It was shown that when a random coding library that was allowed to propagate under optimal growth conditions, it displayed changes in the frequency of individual sequences over time. The authors inferred that sequences that had increased in frequency at the end of experiment provided a fitness benefit to the host relative to all the other sequence variants present. The sequences that increased in frequency did so by allowing its host to grow and reproduce faster than the other competing variants present in the population and therefore they must provide some kind of fitness advantage. Individual effects of the full-length sequences that increased in frequency (referred to as POS_Peps in this thesis) needed to be studied further in order to understand their direct effects on host. Selected POS_Peps allowed for in-depth understanding of individual random peptides and their consequences to the host harbouring them. Since the POS_Peps were enriched from the experiment that was performed under optimal growth conditions (LB media, 37°C growth temperature), none of the sequence variants showed fixation in the populations due to minimal selective pressure. The populations were composed of millions of 195 nucleotides long random sequence variants, each competing for resources. This in itself could have acted as a selective pressure in their experiment (despite avoiding bottlenecks).

In the wild, organisms often face rapidly changing environmental conditions. Temperature is known to be a major driving force that shapes genomes of these organisms (Hochachka and Somero, 2002). Specifically, high temperatures have large effects on the structural and physiological properties of bacteria. Building blocks such as nucleic acids, proteins, lipids and macromolecules (e.g., DNA, RNA, enzymes, proteins and membranes) are affected by high temperature. I hypothesized that if POS_Peps possess the ability to confer fitness advantages to the host, applying a selective pressure may elevate the effects to a more detectable range. Exposing the candidate strains to high and low temperatures served as a relevant starting point to explore beneficial effects of POS_Peps.

This chapter deals with the investigation of three POS_Peps and the effects of their expression on *E. coli* host. POS_Peps were examined to elucidate whether they confer any fitness advantage to the host. Temperature was explored as a selective condition to unravel the presence of conditional advantages of the POS_Peps. A major criticism that has emerged from the previous study was that beneficial effects by the positive frequency variants was an artefact (Knopp and Andersson, 2018). The authors showed an inherent cost of the multicopy vector which when even slightly compensated, increases the frequency and makes the sequences appear to be beneficial. I address these concerns in this chapter and show that the candidate POS_Peps indeed confer a true benefit to the host upon expression.

I have used neutral candidates (referred to as NT_Peps in this study) from the Neme et al. study as controls in addition to the "empty" vector controls (referred as pFLAG in this thesis). Neutral candidates have the same length as the other candidates and a direct comparison between two full-length sequences can be established in this way. NT_Peps were used in order to allow for the comparison of full length products rather than "empty" (pFLAG produces a short peptide ~38 aa) compared to full-length (POS_Pep and NT_Pep = 65 aa). Finally, global gene expression changes in the host as a response to the POS_Peps was studied to gain a better understanding of the functional role of the peptides in question.

5.2. Aims

1. To understand the effects of candidate random peptides on the growth of different *E. coli* backgrounds (K-12 DH10B, B REL606 and B REL607).
2. To perform competitive fitness assays with positive candidate (POS_Pep) strains against the empty vector (pFLAG) controls or neutral candidate peptide strains
3. To study the consequences of POS_Pep expression under variable temperature conditions using competitive fitness assays.
4. To elucidate expression level changes at different time points upon expression of POS_Pep under optimal and high temperature conditions.

5.3. Findings

This chapter describes the effects of random sequences that are neutral in the bacteria and a fraction of sequences that appear to provide an advantage (shown by increase in abundance in the previously described experiment). Sequences were split into three categories: negative (findings described in chapter 3), neutral and positive (described here). The individual effects of three positive candidates (POS_Peps) and their characteristics compared to three neutral candidates (NT_Peps; the candidates that showed no significant change in abundance) along with the empty vector (pFLAG) controls are described in detail.

5.3.1. Selection of candidates for individual exploration

To study the positive impact of random sequences on *E. coli*, individual candidates were selected from the previously described study, where a population expressing random coding library was studied through time to compare the changes in the abundance of each individual sequence (Neme et al., 2017). Three candidates that increased in frequency (Table 5.1, POS_Peps) and three that remained for which, no significant change was observed (Table 5.1, NT_Peps) were chosen for closer comparison (positive versus neutral effects). Selected candidates were cloned on a vector (pFLAG-CTC) and transformed into different *E. coli* strains (K-12 DH10B, B REL606 and B REL607) for further exploration (see methods 2.2.4).

Sequence name	Original ID	Log2 fold change	P-value
POS_Pep1	PEPNR00000000600	4.66	3.9E-38
POS_Pep2	PEPNR00000000032	1.51	1.0E-06
POS_Pep3	PEPNR00000000004	1.23	2.7E-05
NT_Pep1	-	-0.10	0.82
NT_Pep4	-	0.14	0.78
NT_Pep5	-	0.32	0.51

Table 5.1. Positive candidates selected on the basis of fold change and p-values (as analysed with DeSeq2 in Neme et al., 2017).

Candidates that increased in abundance over time were categorised as positive (POS_Peps). Three candidates that were chosen here are the same candidates that were described in the previous study (Neme et al., 2017). The candidates were preliminarily shown to have positive effects in *E. coli* and are further explored here. Candidates that had a non-significant change in their numbers were considered to be neutral (referred as NT_Peps), i.e. they remained constant throughout the experiment and hence had no effect on the host fitness. Three candidates were chosen with the only criterion that the starting normalized count value for each candidate should be above 15. The values reported here are taken from the published dataset in Dryad: http://dx.doi.org/10.5061/dryad.6f356.

5.3.2. Growth rates of strains expressing POS_Peps show no significant differences at 37°C

Candidates expressing the POS_Peps were tested for growth differences with pFLAG controls. Growth measurements were taken (using Tecan Infinite® M Nano^{+} microplate reader) for 24 hours at 37°C with intervals of 30 minutes between consecutive absorbance readings (at 600 nm). Orbital shaking was performed 5 minutes before the absorbance was measured. Three candidates POS_Pep1, POS_Pep2 and POS_Pep3 with pFLAG control against DH10B background are shown (Figure 5.1). Growth curves show differences in the carrying capacity of induced candidates compared to the control pFLAG (Figure 5.1A). I ignore differences in carrying capacities here because the growth data has been collected using a microplate reader, which uses turbidity as a proxy for growth. During stationary phase, since there is a higher density of cells, they tend to form clumps leading to a noisy and unreliable output. Hence, I use growth rates (slopes of curves) as a proxy for fitness. Growth rates showed no significant differences in the three POS_Peps at 37°C (Figure 5.1B). To get an overview of all the p-values of growth rates tested against different backgrounds, refer to Table 5.2. The table summarises different combinations of backgrounds and candidates under induction all compared to the induced pFLAG control. For within strain comparisons i.e. induced and non-induced see Figure 5.1B.

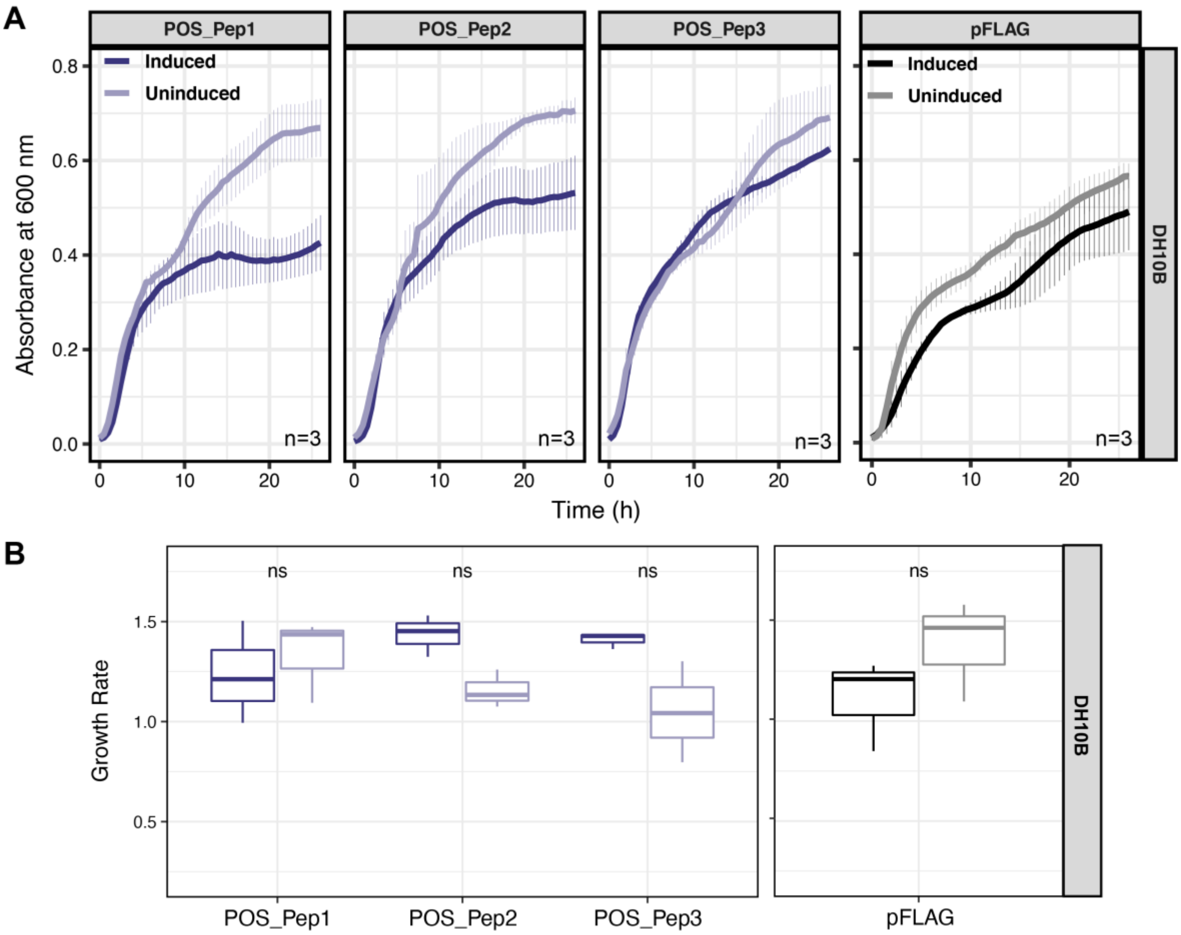

Figure 5.1. Growth dynamics of two positive candidates shows no growth differences in DH10B background.

A) Growth curves show three candidates POS_Pep1, POS_Pep2 and POS_Pep3 (purple), along with pFLAG control (greyscale). Error bars represent standard error of the mean (SEM). Number of replicates per strain = 4. **B)** POS_Peps shows no significant differences in growth rates between induced (dark) and non-induced (light) strains (Wilcox test, p> 0.05) in the DH10B background. P-values calculated using Wilcox test, ns = non-significant. Growth curves performed at 37°C in M9+ Glucose+ Amp [50 µg/mL] media for 24 hours with absorbance recorded at 600 nm every 30 minutes (refer methods 2.2.7.1). Growth rates were calculated using a command line program *GrowthRates* v4.2 (Hall et al., 2014).

Two different genetic backgrounds were tested in order to confirm the effects across strains. Two candidates POS_Pep2 and POS_Pep3 showed no significant differences in the growth rates before and after induction of the POS_Peps at 37°C in REL606 (Figure 5.2A) as well as MG1655 (Figure 5.2B) backgrounds. Growth rates were computed using the maximum slope from the sliding window of five consecutive time points in the log phase using a command line program, GrowthRates v4.2.

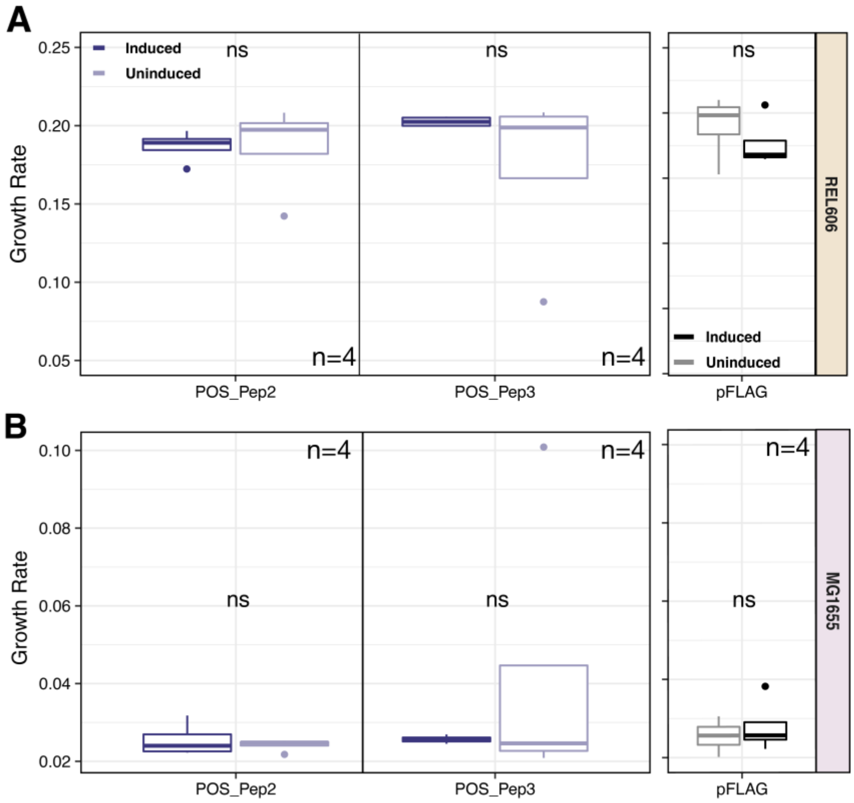

Figure 5.2. Growth rates of POS_Pep2 and 3 tested in two genetic backgrounds REL606 and MG1655 show no significant differences. Growth rates calculated from slopes of the growth curves (log phase) show no significant differences in the induced (dark purple) POS_Peps compared to the uninduced (light purple) controls (Student's t-test, p > 0.05) in **A)** REL606 and **B)** MG1655 backgrounds. In black and grey are the pFLAG controls induced and non-induced respectively. Number of replicates for each strain = 4. Growth rates were calculated using a command line program GrowthRates v4.2 (Hall et al., 2014). Ns = non-significant.

5.3.3. NT_Pep strains show no significant differences as expected

NT_Pep candidates were selected from the library of random sequences that did not change significantly in the course of serial passaging in the previous study (Neme et al., 2017). In other words, these sequences had no effect on their frequencies by the end of the experiment. These candidates were chosen for two reasons: (i) they make products of the same length as the candidates under study (unlike the empty vector controls that make a shorter fragment) and (ii) their neutral phenotype can be used for a direct comparison with any beneficial effects associated with selected POS_Peps. Growth curves of three neutral candidates, NT_Pep1, NT_Pep4 and NT_Pep5 are shown in Figure 5.3. As expected, the candidates showed similar growth trends with (dark) and without (light) IPTG induction (Figure 5.3A). NT_Peps (in gold) show similar growth patterns to the pFLAG controls (in greyscale). Growth measurements were taken with the same parameters as described in the previous section (5.3.2). Growth rates between the two treatments showed no significant differences (p>0.05, Figure 5.3B). Growth rates were computed using the maximum slope from the sliding window of five consecutive time points in the log phase using a command line program, GrowthRates v4.2. As expected, NT_Pep showed no differences in their growth dynamics following expression of the peptides.

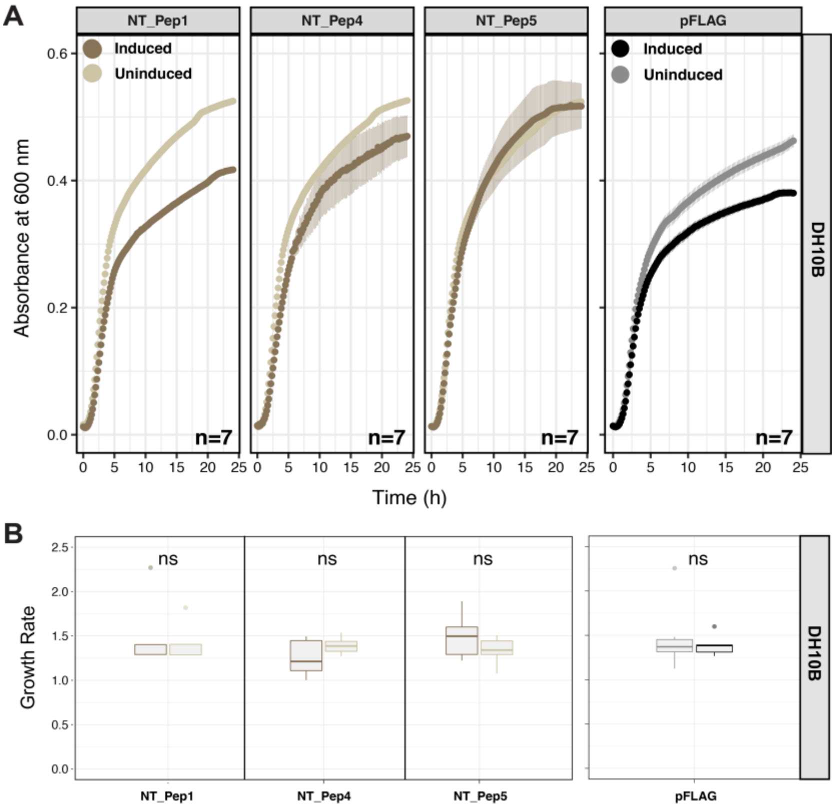

Figure 5.3. Neutral candidates show no significant differences in growth in DH10B background. **A)** Growth measurements of three NT_Peps (coloured) and pFLAG control (greyscale) in DH10B background are shown. For each strain, induced (dark) and non-induced (light) treatments are shown. 'n' is the number of replicates per strain. **B)** Growth rates (calculated from the slope of the curves) are depicted with the same samples in each of the boxes (colours match the top panel). No significant differences (p-values > 0.05) between the growth rates of induced versus non-induced strains (as well as the control pFLAG). All candidates are present in the inducible vector pFLAG-CTC against a K-12 DH10B strain background. Growth curves performed at 37°C for 24 hours and readings were taken every 30 minutes. T-test; P-values: ** < 0.01, *** < 0.001 and ns > 0.05; non-significant. Growth rates were calculated using a command line program GrowthRates v4.2 (Hall et al., 2014).

Background	**DH10B**		**REL606**	
Temperature	**37°C**	**40°C**	**37°C**	**40°C**

POS_Pep1	0.57	0.17	No data	0.4
POS_Pep2	0.12	0.97	0.84	No data
POS_Pep3	0.15	0.33	0.39	No data
NT_Pep1	0.93	No data	No data	0.43
NT_Pep4	0.216	No data	No data	0.88
NT_Pep5	0.93	No data	No data	0.79

Table 5.2. Summary of p-values for induced POS_Peps and NT_Peps compared to the corresponding pFLAG control.

P-values calculated by Student's or Wilcox t-test show non-significant differences ($p>0.05$) for all tested combinations in two backgrounds DH10B and REL606. Two backgrounds and two temperatures (37 and 40°C) are shown here. No data represent the combinations that were not tested. All growth curves were performed in 96 welled plates and absorbance read at 600 nm using a microplate reader. Growth rates were calculated using a command line program GrowthRates v4.2 (Hall et al., 2014).

5.3.4. Candidate POS_Pep1 shows competitive fitness advantage at 40°C (sub-optimal temperature)

Measurement of maximum growth rates of cultures using turbidity is widely used for real time analysis of bacterial populations. Although the method is simple and rapid, it fails to capture subtle differences in the population dynamics and only provides a proxy for fitness. In order to increase the sensitivity of detection and to provide a closest estimate for fitness, competitive fitness assays are used. The competitive fitness is calculated simply by counting the number of colonies (CFU/mL) of each competitor at the beginning (T_0) and the end (T_{24}) of competition. Genotypes with higher fitness tend to produce more offspring and out compete their less fit competitors.

For competition assay, I used a previously established methodology with two ancestral strains *E. coli* B REL606 and REL607, which produce red and pink coloured colonies on tetrazolium arabinose (TA) agar plates respectively ((Lenski et al., 1991), described in 2.2.9). In brief, REL606 differs from REL607 by a single point mutation in the *araA* gene, which makes it an auxotroph for L-arabinose metabolism. The strains that can utilize L-arabinose on the TA plates acidify the surroundings turning the colony colour from red to white, while the non-utilizing strain REL606 produces red coloured colonies. For this, NT_Pep expressing strains can be used as additional controls (pFLAG being the empty vector control) to capture the subtle fitness effects that POS_Pep might provide. To investigate the effects of expressing POS_Pep1 on host, competitive fitness estimation against pFLAG containing strains was performed. The assay was performed at three different temperatures- optimum (37°C), low (33°C) and high (40°C) to test selective effects, if any.

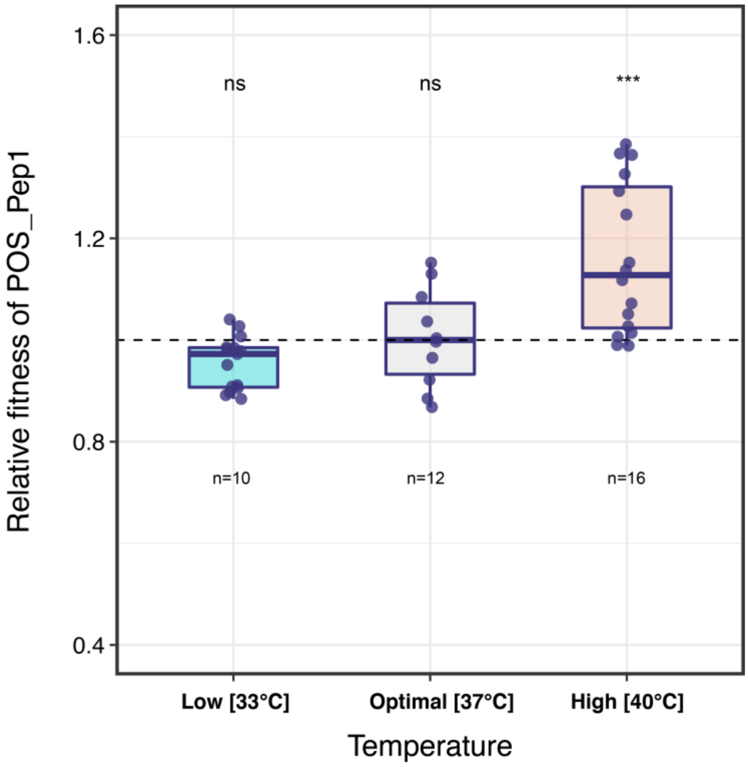

Figure 5.4. POS_Pep1 shows competitive advantage over control (pFLAG) at 40°C. The box plot shows relative fitness of POS_Pep1 compared to pFLAG control measured at the end of 24 hours at three different temperatures (33°C, 37°C and 40°C) with IPTG induction. No significant difference (p > 0.05) seen in the at 37°C (light grey) and at 33°C (blue) the POS_Pep1 loses competition with a decreased relative fitness (p=0.005; no increase in fitness) after 24 hours. At 40°C, POS_Pep1 strains show a significantly higher relative fitness (p = 0.0007) compared to pFLAG controls (orange) after 24 hours. The dashed line at relative fitness value of 1 represents no difference between the two competitors (here POS_Pep1 and pFLAG) and any deviation from this line represents higher or lower relative fitness. pFLAG represents the empty vector control and POS_Pep1 represents the candidate random sequence. Assay was performed with candidate vectors cloned in either REL606 (red colonies) or REL607 (pink colonies) backgrounds (as well as in swapped backgrounds). Data points shown are relative fitness of strain A/B pooled from both backgrounds. 'n' is the number of individual replicates. Competitions were performed in M9 glucose media (Amp + IPTG) under shaking conditions. One sample t-test p-values: ns=non-significant, *** <0.001.

In the competitive fitness assays, each competitor was mixed in a one to one ratio from a fully-grown overnight culture (~10^9 cells/mL) and allowed to compete in IPTG induced minimal medium (M9+Gucose) for 24 hours at three temperatures. Note that competitive fitness assays measure the colony numbers as opposed to the growth rates calculated from the slope of growth curves, which measures the turbidity of the medium. Competitive fitness measurements are far more sensitive than the growth curves, hence preferred when small differences need to be observed. Cell densities of each population types were counted at initial (T0) and final (T24) time points and relative fitness was calculated using previously described Malthusian parameter (see 2.2.9.1). Swapped background strains were also engineered and tested to eliminate background related differences; the data points in box plots are pooled values (REL606 plus REL607 background). POS_Pep1 (versus pFLAG) did not show significant difference ($p > 0.5$) in its competitive fitness at the optimal temperature of 37°C (Figure 5.4A). Intriguingly, POS_Pep1 (competed against pFLAG) showed a significantly increased (p=1.4e-06) relative fitness when competed at an elevated temperature (40°C, Figure 5.4B, orange fill), but not at a low temperature (33°C, p>0.05, Figure 5.4B blue fill). Candidate POS_Pep1 is advantageous to the host at high temperature, perhaps by providing certain conditional benefits.

5.3.5. Candidate Pos_Pep2 displays analogous competitive fitness advantage to POS_Pep1 at 40°C

Positive peptide candidate POS_Pep1 showed an increased fitness at 40°C when competed against the empty pFLAG control. To investigate a similar selective advantage at high temperature, another candidate POS_Pep2 was tested for its competitive fitness against pFLAG control. Competition assay was performed at 37°C and 40°C for 24 hours. Relative fitness values of POS_Pep2 compared to pFLAG under optimal temperature conditions (37°C) show no significant differences (p=0.29; Figure 5.5; 37°C). POS_Pep2 shows a significantly higher relative fitness p=0.005) compared to the pFLAG control at 40°C under IPTG induction (Figure 5.5; 40°C). POS_Pep2 similar to POS_Pep1 shows clear and reproducible fitness advantage selectively at a higher temperature. Under IPTG induction the background express the POS_Peps as well as pFLAG "empty" control (that produces 38 aa product). Under optimal temperature conditions both appear to grow at the same rate (hence relative fitness close to 1) but POS_Peps provide a fitness advantage at 40°C such that the hosts expressing them can easily outcompete the pFLAG controls. Candidate random peptides have the ability to confer selective fitness advantage to the host upon expression.

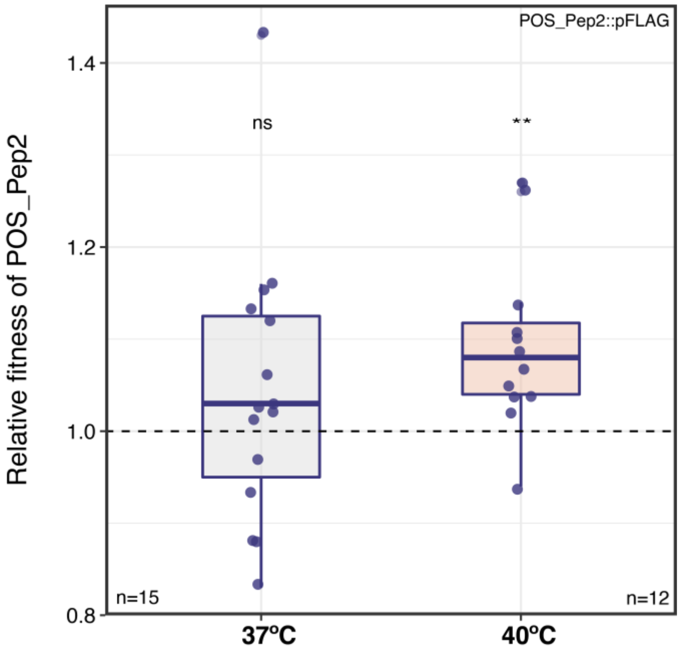

Figure 5.5. POS_Pep2 shows fitness advantage only at high temperature (40°C). The relative fitness values of POS_Pep2 (grey) relative to pFLAG control measured at the end of 24 hours at 37°C with IPTG induction. Relative fitness values show a high variation with no significance (p = 0.29). At 40°C, POS_Pep2 (orange) shows significantly higher relative fitness (p = 0.005) compared to pFLAG controls at the end of 24 hours. Competition assay was performed with candidate vectors cloned in either REL606 (red colonies) or REL607 (pink colonies) backgrounds (as well as in swapped backgrounds). Data points shown are relative fitness of strain A/B pooled from both backgrounds. Competitions were performed in M9 glucose media (Amp + IPTG) under shaking conditions (2.2.7.1). One sample t-test, p-values: ns (non-significant)>0.05 and **<0.01.

5.3.6. POS_Pep1 has a fitness advantage when competed with NT_Pep4

Although competing the POS_Peps with pFLAG control provided strong evidence for possible advantageous function of these peptides under high temperature stress, pFLAG does produce a small peptide even though it does not harbour any insert sequence (e.g. the intact multiple cloning site is present downstream of the promoter). Therefore, competing against a candidate that produces a full length product (of 65 aa), same as the candidate in question, would indeed provide a stronger support to the previous findings (Figure 5.4 and Figure 5.5).

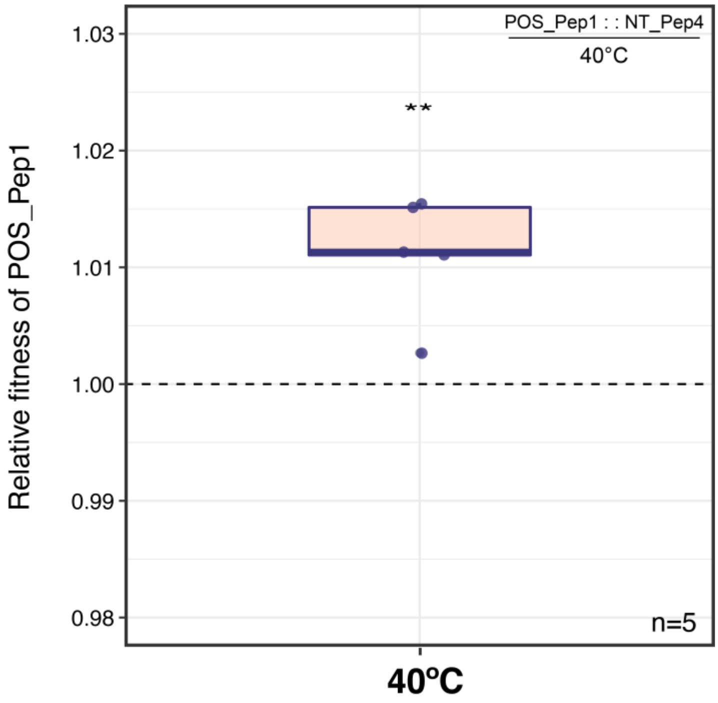

Figure 5.6. POS_Pep1 provides a competitive fitness advantage against the neutral sequence NT_Pep4 at 40°C. The plot shows relative fitness values of POS_Pep1 (orange) against NT_Pep4 measured at the end of 24 hours at 40°C with IPTG induction. POS_Pep1 has a significantly higher relative fitness (p = 0.0085). Assay performed with POS_Pep1 in strain background REL606 (red colonies) and NT_Pep4 in REL607 (pink colonies) under shaking conditions. One sample t-test; p-value ***<0.001.

POS_Pep1 (in REL606 background) was competed with NT_Pep4 (in REL607 background) at high temperature (40°C) and the relative fitness was calculated at the end of 24 hours. POS_Pep1 showed a significantly higher relative fitness (p=0.0001) compared to the NT_Pep4. A swapped background (i.e. REL607-POS_Pep1 versus REL606-NT_Pep4) competition is not included in the data points which is essential to eliminate effects from the background. These results were preliminary and swapped background confirmation will have to be done to provide a stronger evidence. Although the swapped control was not performed, previous experiments (Figure 5.4 and Figure 5.5) have repeatedly shown that the REL606 and REL607 background have no effects and serve as neutral ancestors for competitions. Representative POS_Pep1 retains its beneficial effects even when competed against a NT_Pep, suggesting a peptide mediated advantage conferred to the host at high temperatures.

5.3.7. Beneficial candidates do not show enhanced expression of stress response genes

A microarray experiment similar to the one in chapter 3 was performed, to monitor the changes in the mRNA expression (using microarray *E. coli* K-12 chip G4813A-020097) before and after induction of expression of beneficial candidates. The experiment was performed with POS_Pep1, POS_Pep2 and pFLAG control against REL606 background. The genetic background (REL606), media (M9+ Glucose) and temperature (40°C) conditions used were same as the competition experiments, where a fitness advantage in the two POS_Peps was observed. Total RNA extracted at three different time points was used for hybridization to the chip (see methods). Time point 1 consisted of non-induced culture from the exponential phase. The subsequent time point 2 consisted of culture induced for 1 hour and time point 3 consisted of culture induced for 16 hours. For each temperature, there were five replicates for every time point - two candidates (POS_Pep1 and POS_Pep2) and three replicates for empty vector controls.

Several genes were differentially expressed in the two POS_Pep expressing candidates (POS_Pep1 and 2) when compared with the pFLAG control under similar conditions (Figure 5.7). Under high temperature, the differential expression of POS_Pep1 after one hour induction (Figure 5.7A) showed that 66 genes had enhanced expression and 48 genes had decreased expression in cells expressing POS_Pep1. After 16 hours of induction, 667 genes had enhanced whereas 318 genes had decreased expression in the POS_Pep1 expressing cells. Similarly, for POS_Pep2, 1204 genes showed increased expression and 1643 showed decreased expression compared to pFLAG after one hour of IPTG induction (Figure 5.7B). After 16 hours of induction, 840 genes show increase whereas 421 genes show significant decrease in expression. Larger number of genes show differential expression in POS_Pep2 compared to POS_Pep1, both compared with pFLAG. This demonstrates that gene expression patterns are indeed affected by peptide expression under the higher temperature.

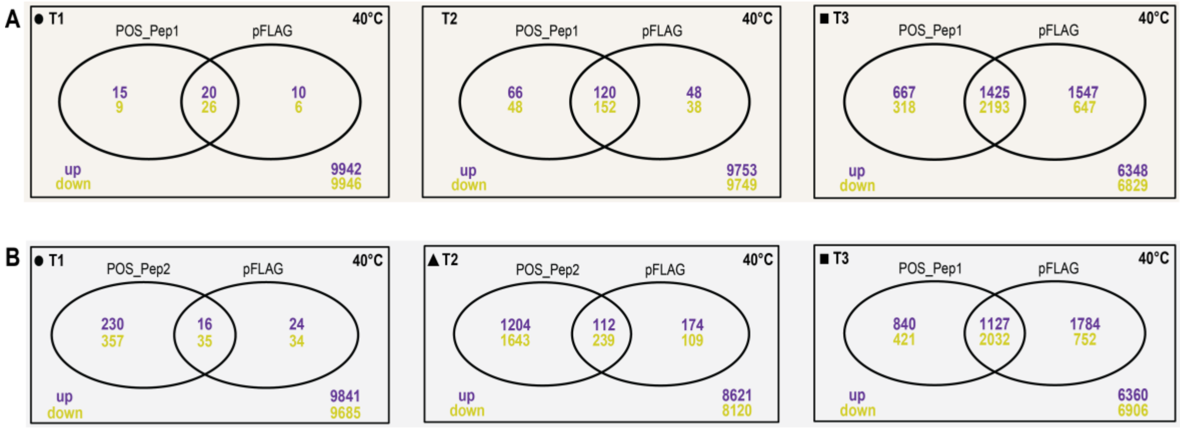

Figure 5.7. Venn Diagram showing differentially expressed genes at the three sampled time points in POS_Pep1 and pFLAG control grown at 40°C.

Venn diagrams for POS_Pep1 and 2 compared to pFLAG show differentially expressed gene numbers from time points T1, T2 and T3 at high temperature (40°C). **A)** Comparison of differential expression in POS_Pep1 compared to pFLAG. **B)** Comparison of differential expression in POS_Pep2 compared to pFLAG. Differentially expressed genes in samples from the exponential phase (T1; non-induced), followed by after 1 hour of induction (T2; induced), when cells are still in the exponential phase. Differentially expressed genes after 16 hours of induction when the cells are in the late stationary phase (T3; induced). Numbers in purple and yellow are genes that show higher and lower expression respectively and the numbers outside the ellipses are genes that had non-significant differential expression (p>0.05).

Principle component analyses (PCA) of POS_Pep 1 and 2 shows majority of variation contributed by time point 3 (T3- 16 hours IPTG), which is the expression profile of samples in the late stationary phase after induction (Figure 5.8). POS_Pep1 shows close cluster of T1 and T2, suggesting that the expression of these samples does not differ much (Figure 5.8A). POS_Pep2 shows a clearer clustering of time points T1 and T2 on the PC2 axis that is explains 18.2% variance in the samples (Figure 5.8B). PC1 axis explains 37.5% variance, which clusters T3 from the T1 and T2. A similar distribution is seen in PCA for pFLAG at 40°C across three time points (Figure 5.8C). No clear distinction observed between pFLAG and POS_Peps at 40°C suggesting presence of subtle effects, if any.

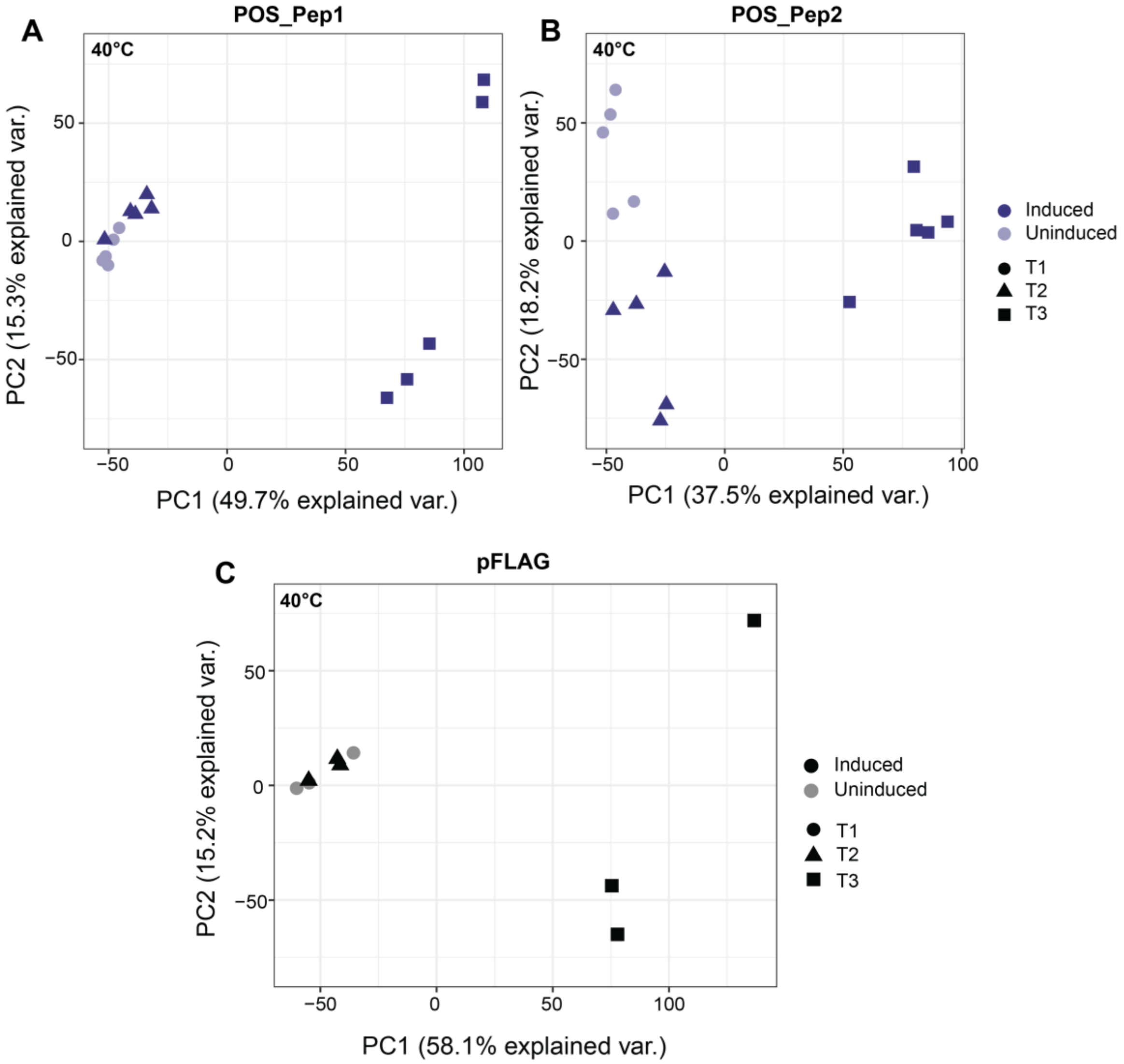

Figure 5.8. PCA of POS_Pep1 and 2 of expression data sampled at T1, T2 and T3 at 40°C in REL606 background.

Three time points are represented as T1- non-induced (circles), T2- one hour post induction (triangles) and T3- 16 hours post induction. **A)** PCA for POS_Pep1 and **B)** PCA for POS_Pep2 and **C)** for pFLAG, at 40°C. Clear distinction can be seen along the three time points, especially at T3 but no visible differences between pFLAG and POS_Peps.

5.3.8. Genes differentially expressed at high temperature appear to have involvement in transport related functions

In chapter 3, the effects of NEG_Peps on *E. coli* host were discussed in detail. It was shown that after one hour of IPTG induction, NEG_Pep expressing strains had an increased expression of genes involved in the general stress response pathways. This general stress response was likely a coping mechanism in order to tolerate the accumulating deleterious peptides. Contrary to that observation, in POS_Pep strains, the stress response genes did not show significant differential expression upon induction (Figure 5.9; also see stress gene expression in NEG_Pep5 chapter 3, Figure 3.7). In POS_Pep1 and 2, no evidence of increased expression of genes related to the general stress response was observed, which indicates that both beneficial POS_Peps were well tolerated within the host (see Table 5.3). Several genes that had increased expression in NEG_Pep5 fell in the category of stress response genes in the gene ontology analysis. These candidates were selected and their corresponding values in POS_Pep1, 2 and pFLAG were checked for high temperature expression data sets. None of the genes had the log2 fold change values above 2 and the majority of them did not even surpass a log2 fold change of 1. All the listed stress genes showed a log2 fold change value > 1.5 in the NEG_Pep5. In POS_Pep1 and 2 these genes did not show significant changes and remained in the fold change threshold of 0-1 (Table 5.3). Although a clear explanation about the fitness advantage conferred to host at 40°C could not be established at this point, a large number of genes showed differential expression after induction, which could probably be due to a global cellular response. Absence of increased expression of the stress related genes indicates that induction of POS_Pep1 and 2 expression does not have a deleterious effect on their host.

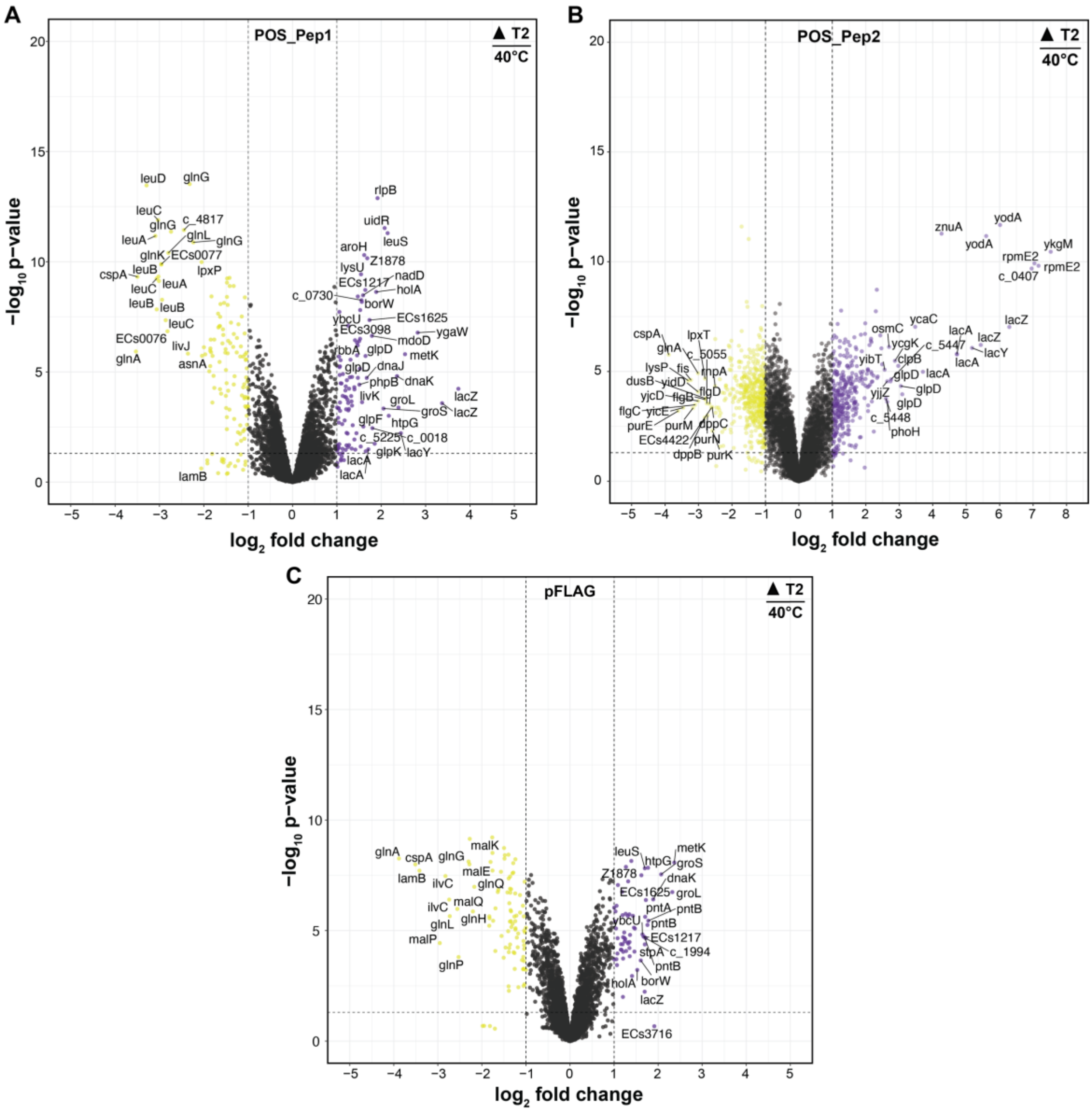

Figure 5.9. No evidence of enhanced expression of stress response genes in POS_Pep1 post induction.

Volcano plots depict the differential expression of genes at 40°C after 1 hour of IPTG induction (T2) normalized to their respective non-induced time point 1. **A)** POS_Pep1 shows several genes that were differentially expressed upon peptide induction, but stress genes (as seen in the NEG_Pep5, chapter 3) seem to be absent. **B)** POS_Pep2 shows several genes differentially expressed after induction compared to pFLAG. **C)** Control (pFLAG) expression post induction shows differentially expressed genes at 40°C. POS_Pep1, POS_Pep2 and pFLAG control were both in the B REL606 background, growing on M9+Glucose media. Shaking cultures (40°C, 250 RPM) were sampled in exponential phase (T1) followed by induction for 1 hour using IPTG (1 mM) before sampling T2. Yellow dots represent genes that decreased in fold change and purple dots are the genes that increased in fold change compared to respective T1 values. Vertical and horizontal dotted lines are the manually set cut-off values for fold change (± 1)and p-values (0.05). Expression data analysed with R version (3.6.3 and 4.0.0) using the Limma Bioconductor (v3.11) package (Smyth, 2005).

Stress genes	**POS_Pep1**	**POS_Pep2**	**pFLAG**
	T2 [log2 FC]	**T2 [log2 FC]**	**T2 [log2 FC]**
bhsA	0.08	1.40	0.18
degP	0.81	0.73	0.64
grpE	0.74	0.69	0.58
hspQ	0.18	0.88	-0.21
htpX	0.6	-0.29	0.34
ibpA	0.55	1.11	0.15
ibpB	0.51	1.82	0.19
lon	0.01	0.06	-0.26
marA	0.32	-0.48	0.35
marB	0.11	-0.35	0.03
marR	0.29	-0.95	0.32
pspA	0.48	1.09	0.37
pspB	0.26	0.68	0.15
pspC	0.34	0.63	-0.07
pspD	0.34	0.48	-0.07
pspG	-0.08	-0.39	-0.1
pstS	0.51	0.09	0.6
relB	0.01	0.05	0
relE	0.06	0.04	-0.1
rseA	0.01	0.88	-0.11
soxS	0.79	0.77	0.26
ybbN	0.05	0.37	-0.17
ybiJ	0.29	1.60	-0.16

Table 5.3. Log2 fold change values for stress related genes that had increased expression in NEG_Pep5 (see chapter 3, Figure 3.2).

Log2 fold change of stress related genes in POS_Pep1 and POS_Pep2 after one hour of induction 40°C together with pFLAG. This log2 fold change threshold is set arbitrarily but used in the context of NEG_Pep5 values which were at least above 1.5 for the candidate genes shown. The three candidates shown were each compared to their respective non-induced controls and the genes that showed variation between the two experimental conditions were eliminated (by comparing pFLAG of the two experiments). The columns with POS_Pep1, 2 and pFLAG comprise of the log2 fold change values from the expression data analysis performed using Limma package in R (methods 2.2.12.4). Adjusted p values are shown in the last column.

5.3.8.1. Gene ontology (GO) analyses to determine functional classes in POS_Pep1 at 40°C

To understand the gene classes further, a gene ontology (GO) analysis was performed on the top 100 highest fold change candidate genes. Table 5.4 shows the enriched functional categories for top 100 candidate genes from POS_Pep1 and empty vector control for T2 (one hour induction) at 40°C. POS_Pep1 candidate strains showed multiple functional categories related to transport and localization of molecules (highlighted orange Table 5.4). None of the other controls showed these categories. They were also absent in the same candidate (POS_Pep1- T2) at 37°C. Elevated expression of genes involved in transport and localization was found to be a unique characteristic of strains expressing beneficial peptides at high temperatures. Two genes, *yghG* and *pppA* showed enhanced expression only in the POS_Pep1 post induction at 40°C. Both these genes are known to be required for the assembly of Type II secretion system in B which are specific to these strains.

Although the cellular function of beneficial random peptides could not be determined, a clear phenotype was observed in the competitive fitness measurements. Therefore, random peptides (POS_Peps) have the potential to display beneficial effects under adaptive conditions. This chapter demonstrates that not all random over-expressing peptides are deleterious to the host (Weisman and Eddy, 2017) but it depends on the properties of peptide and the ability of host to adapt to them.

POS_Pep1 **T2**				**pFLAG** **T2**			
Enrichment FDR	**Genes in list**	**Total genes**	**Functional Category**	**Enrichment FDR**	**Genes in list**	**Total genes**	**Functional Category**
4.50E-34	54	3215	biological process	8.00E-24	45	3215	biological process
1.10E-25	45	2611	cellular process	1.10E-22	41	2611	cellular process
2.70E-24	41	2142	single-organism process	3.00E-21	39	2494	metabolic process

2.30E-23	38	1816	single-organism cellular process	3.00E-21	35	1816	single-organism cellular process
1.20E-19	39	2494	metabolic process	5.60E-19	35	2142	single-organism process
7.00E-19	35	1957	organic substance metabolic process	4.30E-18	33	1950	cellular metabolic process
2.20E-18	33	1733	primary metabolic process	4.30E-18	33	1957	organic substance metabolic process
7.90E-16	32	1950	cellular metabolic process	2.20E-16	30	1733	primary metabolic process
7.90E-13	23	1097	biosynthetic process	1.10E-14	26	1376	single-organism metabolic process
9.40E-13	25	1376	single-organism metabolic process	2.50E-12	19	766	small molecule metabolic process
2.90E-11	21	1049	cellular biosynthetic process	1.30E-11	21	1097	biosynthetic process
3.40E-11	21	1062	organic substance biosynthetic process	5.30E-11	20	1049	cellular biosynthetic process
1.00E-10	12	238	cellular amino acid metabolic process	4.60E-10	21	1336	nitrogen compound metabolic process
1.00E-10	18	766	small molecule metabolic process	5.10E-10	15	570	single-organism biosynthetic process

7.00E-10	14	444	carboxylic acid metabolic process	4.80E-09	18	1062	organic substance biosynthetic process
8.50E-10	5	10	branched-chain amino acid transport	5.20E-09	15	679	organonitrogen compound metabolic process
1.00E-09	11	227	ion transmembrane transport	2.00E-08	17	1024	organic cyclic compound metabolic process
1.10E-09	14	466	oxoacid metabolic process	6.90E-08	12	469	organonitrogen compound biosynthetic process
1.10E-09	14	468	organic acid metabolic process	9.30E-08	16	991	cellular aromatic compound metabolic process
1.20E-09	15	570	single-organism biosynthetic process	2.30E-07	9	238	cellular amino acid metabolic process
1.50E-09	13	393	organic substance transport	3.60E-07	6	67	sulphur compound biosynthetic process
1.50E-09	8	82	organic acid transmembrane transport	4.30E-07	16	1116	cellular nitrogen compound metabolic process
1.80E-09	16	700	transport	5.20E-07	15	982	heterocycle metabolic process
1.90E-09	16	706	establishment of localization	1.90E-06	3	5	hydrogen sulphide

							metabolic process
3.50E-09	16	738	localization	1.90E-06	3	5	hydrogen sulphide biosynthetic process
4.70E-09	7	59	amino acid transmembrane transport	2.20E-06	8	231	nucleobase-containing small molecule metabolic process
8.60E-09	20	1336	nitrogen compound metabolic process	2.30E-06	6	95	sulphur compound metabolic process
9.60E-09	15	679	organonitrogen compound metabolic process	2.90E-06	10	442	phosphate-containing compound metabolic process
9.60E-09	13	469	organonitrogen compound biosynthetic process	2.90E-06	11	561	aromatic compound biosynthetic process
1.30E-08	8	112	carboxylic acid transport	2.90E-06	10	444	carboxylic acid metabolic process

Table 5.4. Gene ontology (GO) terms from top 100 genes shows enrichment of transport related genes in POS_Pep1, T1 (1 hour induction) at 40°C.

Top 100 high fold change genes taken from T2 of POS_Pep1 shows increase in expression of several genes that have known transport related biological functions (highlighted in orange) which do not show enrichment in controls as well as in the expression data from 37°C (not shown here). GO terms were computed from top 100 high fold change genes from the 40°C expression set at T2 (1 hour of induction) in candidate POS_Pep1 and control pFLAG. Total genes are the number of genes in a particular functional category from the reference genome, *E. coli* K-12 DH10B. GO term analysis was done using the web-based software ShinyGO v0.61: Gene Ontology Enrichment Analysis (Ge et al., 2019).

5.4. Discussion

Random peptides with the potential to produce deleterious effects in *E. coli* have been described in chapter 3. The goal of this chapter was to understand the role of random peptides that had previously been shown to increase in frequency (POS_Peps) when expressed as a part of random sequence library on a vector and allowed to grow for several generations in *E. coli* host (Neme et al., 2017). About 25% sequences that showed increased frequency amid the pool of millions of other random sequences were reported to have a fitness advantage over the other candidates. Individual POS_Pep candidates that have been described in this chapter were chosen from this previous study.

It was found that POS_Peps confer a conditional fitness advantage. The fitness advantage was found to be robust and reproducible when the bacteria were grown at 40°C as opposed to the optimal temperature 37°C used for enrichment in the previous study of Neme et.al. Three POS_Peps were chosen for the study together with three NT_Peps (sequences that did not change significantly in frequency in the previous study) as controls (along with the pFLAG control). At optimal growth temperatures (37°C) the strains expressing POS_Peps had no discernible differences in growth rates compared to the pFLAG controls or to the NT_Peps. Competitive fitness assays at 37°C also showed no statistical differences in the relative fitness values of POS_Pep strains compared to pFLAG. Since the growth conditions were not restrictive (37°C) it is likely that the fitness effects are not distinguishable. A previously described study had shown that selection of randomized DNA libraries under stress improves the fitness of the peptide-expressing host strains (Stepanov and Fox, 2007). Another study showed an enrichment of beneficial random peptides from a random sequence library in the presence of antibiotic selective pressure (Knopp et al., 2019). Under selective conditions new phenotypes can spread faster and can be more readily distinguished. The POS_Peps studied here were from the sequences that had shown to be enriched after passaging for four successive cycles, under relaxed conditions (optimal growth conditions; LB media and 37°C) with minimal selective pressure. Moreover, the beneficial effects of individual candidates on the host could be rather subtle when tested under similar optimal conditions without serial propagation. Exposing the candidates to a stressful environment may allow for better visualization of the underlying beneficial effects. In the wild, a commonly observed physical environmental factor is changing temperatures, which was used as a starting point to determine if temperature changes can highlight the fitness differences in POS_Pep strains compared to pFLAG control. Indeed, at 40°C POS_Pep expressing strains showed a fitness advantage compared to the pFLAG control. High temperature indeed gives a conditional advantage to the strains expressing the POS_Peps. The beneficial effect of the POS_Peps persisted even when competing against the NT_Pep expressing strains. This indicates that the host not only tolerates the over-production of the POS_Peps, but also displays no negative effects on the global gene expression. Similar results were obtained at optimal and high temperatures for two POS_Pep expressing strains. This shows that different peptides are recognized differently in the host; some are deleterious while others are not. In other words, not all over-expressing peptides have a deleterious effect on the host.

While the stress response genes did not show elevated expression levels in POS_Peps, the mechanism behind the beneficial effect at high temperature (40°C) remains unclear. The expression profiles of the induced POS_Pep1 strains compared with the pFLAG controls revealed a number of differentially expressed genes (see Figure 5.7). GO term analysis of the top 100 genes in POS_Pep1 after induction showed several clusters of genes involved in transport and localization that were absent in the corresponding pFLAG controls (see Table 5.4). Efficient and fast utilization of resources using the active transport could be a candidate mechanism by which POS_Pep1 strains outcompete the controls. Beneficial effects of positive random peptides (POS_Peps) are readily observed phenotypically, but the underlying mechanisms of how these peptides interact with the host machinery needs further investigation.

Chapter 6. Discussion

6.1. Review of findings

A proof of concept study to understand the process of *de novo* gene evolution can be accomplished by testing for specific biological activity of random stretches of nucleotides and amino acids. The question of what percentage of random sequences can be bioactive was answered by Neme and colleagues from our lab, in 2017. Although the study suggested a biological functionality of random sequences based on their changed frequencies, the impact on the host remained unclear. This begged the question- what are the effects of individual random peptides on the host fitness?

6.1.1. Bioactivity of random peptides

My investigation of individual random sequences and their effects in *E. coli* confirmed that random peptides can possess bioactivity. Individual candidates from sequences that decreased (NEG_Peps), increased (POS_Peps) and remained unchanged (NT_Peps) were selected for this study. Each candidate was characterized with regards to their fitness effect on the host via growth rate measurements as well as direct competition experiments. In this thesis, I demonstrate that the candidates from each of the three groups and their effects on the host. Over-production of NEG_Peps is deleterious to the host whereas over-expression of POS_Peps, on the other hand, can provide beneficial effects at an elevated temperature. POS_Peps and NT_Peps, the candidates that had higher or non-significant changes in frequencies in Neme et al. 2017 study, show no disadvantage when over-expressed in the host. It has been shown that codon composition of sequences affects translation of proteins (Ermolaeva, 2001). It has also been shown that the speed of translation and protein folding are affected if higher non-optimized codons are present in a sequence (Yu et al., 2015). This could imply that NEG_Peps may have a codon composition that is completely non-optimal, while POS_Peps could have a more optimal set of codons. But this was not the case for any of the candidates in this study. Optimality of codons can be calculated with a codon adaptation index (CAI) metric, which calculates how optimal given sequences are to the provided reference genomes codon usage (Sharp and Li, 1987; Bulmer, 1990). A CAI value of 1 means that the codon usage of the provided sequence is optimal. The data show that none of the sequences were well adapted (CAI ~ 0.5) to the codon usage of *E. coli* (Figure 6.1), which is the expectation in random non-biological sequences.

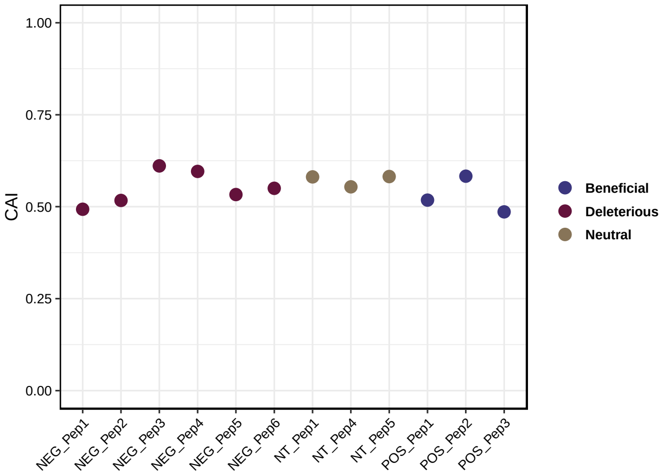

Figure 6.1. Codon adaptation index (CAI) of candidates. The codon adaptation index calculated to determine synonymous codon usage bias for all candidates used in this thesis. Three categories: deleterious (maroon), neutral (yellow) and beneficial (purple) are shown. CAI was calculated using EMBOSS CAI tool (Bulmer, 1990).

Although POS_Peps do not show discernible beneficial effects at optimal growth conditions, the fitness advantage is clearly detected at high temperatures. This is the first evidence of random peptides conferring fitness benefits under stressful high temperature conditions. *E. coli* expressing POS_Peps have a conditional fitness benefit (discussed further in section 6.1.1.2). Together the data suggest, that candidate random peptides can show bioactivity upon expression, which in turn reflects on the growth of the host. The results of chapter 3 and 5 highlight the deleterious and beneficial effects that random peptides have on the host fitness. This leads to the question- how do NEG_Peps and POS_Peps affect the host fitness?

6.1.1.1. NEG_Pep mediated fitness disadvantage in host

In chapter 3, I investigated individual effects of candidate random sequences that decreased in frequency in the previous study (Neme et al., 2017). NEG_Peps were ligated inside an inducible, multicopy vector, pFLAG-CTC (referred as pFLAG) and engineered into three different *E. coli* backgrounds. NEG_Peps conferred a fitness disadvantage to the *E. coli* host due to a prolonged growth lag. What might cause the fitness disadvantage? Induction of NEG_Pep expression leads to a transient arrest in the cell growth, visible as the prolonged lag phase in the growth curves. Whether there is cell death during this prolonged lag will need further investigation. At the end of the lag phase, the growth rate resumes and is comparable to the controls; this suggests that either the cells become tolerant to the peptides through activation of several cellular responses or they behave like persisters (Vulin et al., 2018) by arresting their growth until the stress, here IPTG, in the medium exhausts, before starting to grow normally. This raised questions about what might be happening inside the cells when they are challenged by the deleterious peptides. In the upcoming section 6.1.2, I discuss the differential expression of genes in the strains expressing NEG_Peps and POS_Peps together with their potential implications on the host fitness.

6.1.1.2. POS_Peps provide conditional fitness advantage to hosts

In chapter 5, I investigated the fitness effects of three candidate POS_Peps, that were selected from the Neme et al. 2017 study. POS_Peps showed beneficial effects when competed with the pFLAG controls at elevated temperature (40°C). Fitness differences were hard to disentangle at 37°C, although an upward trend in fitness of POS_Peps was evident. However, the results were variable (low reproducibility) and statistically non-significant most times. Previous studies have shown that applying selection conditions allows the identification of functional random peptides using phenotypic rescue as a proxy (Stepanov and Fox, 2007; Knopp et al., 2019). Raising the growth temperature to suboptimal conditions constitutes a similar approach: I hypothesized that if the POS_Peps provide a generalized fitness advantage under selective conditions, this advantage should amplify detectably. Competing POS_Peps with respective controls at 40°C, indeed provided a fitness benefit, in which POS_Peps managed to take over its competitor in 24 hours. The results with 40°C unlike 37°C were highly reproducible. The same POS_Peps had no significant advantage at lowered temperature (33°C). Together the results suggest that POS_Peps confer a fitness advantage to the host, but only under suboptimal growth conditions. This conditional fitness benefit was not only against the “empty” pFLAG controls but also against the NT_Pep4 (only one neutral candidate tested), which produces a full-length peptide. These findings provided the first evidence of beneficial effects of random peptides in *E. coli*.

6.1.2. Variation in *E. coli* transcriptomes triggered by random peptides

Organisms have the ability to modulate cellular pathways in response to environmental cues, by fine-tuning their gene expression (Bradshaw, 1965; Schlichting and Smith, 2002). Fine-tuning expression instantaneously allows organisms to adapt and subsequently increase chances of survival and reproduction (reviewed in (de Nadal et al., 2011)). I hypothesized that, in order to undergo rapid adaptation under the expression of deleterious NEG_Peps, it would be crucial for the host to maximize survival. This was reflected in several genes that showed enhanced expression (samples taken at the end of one hour of induction) in NEG_Pep expressing strains as compared to the controls (see chapter 3, Figure 3.7). The toxin *hokB*, was seen to be upregulated in the NEG_Pep1 and 5, which has been shown to be a part of the type I toxin-antitoxin system in *E. coli* (reviewed in (Page and Peti, 2016)). It has been shown that elevated levels of HokB cause membranes to depolarize leading to loss of stability (Gerdes et al., 1986). It was also shown that HokB increases persistence following loss of membrane potential in *E. coli* (Verstraeten et al., 2015). This effect potentially explains the initial lag in growth in the cells expressing NEG_Peps followed by a sharp rise after a while. The expression of NEG_Peps possibly causes the membranes to depolarize through the enhanced expression of the HokB toxin. Among other genes that showed increased expression were the ones involved in a general stress response in *E. coli* (see discussion in Chapter 1).

Expression of POS_Peps induce changes in the host transcriptome that were distinct from NEG_Peps. No upregulation of stress response genes at 37°C or 40°C (high temperature tested additionally) was observed. The absence of toxin or stress gene expression in the POS_Peps are an indication that they are not perceived as stress within the host. The fitness advantage provided by POS_Pep was conditional and observed only at 40°C. This begged the question, whether there were any changes in expression specific to 40°C in the host expressing POS_Peps. Although several genes have enhanced expression specifically at 40°C, it was difficult to predict which ones were likely causing the fitness advantage. GO term analysis showed that specifically at 40°C, several genes with increased expression fell into the transport and localization categories. My current hypothesis is that a higher fitness observed at 40°C is a result of interactions of the POS_Peps with genes to increase the capacity to metabolize various nutrients (for example, by efficient transport of nutrients from the outside media) causing them to grow faster than their competitors. Further experiments will be necessary to understand the cellular changes elicited in the host in response to POS_Peps.

6.1.3. Diverse adaptive paths explored by NEG_Pep suppressor genotypes

6.1.3.1. Potential targets of the NEG_Peps in the host

Due to the deleterious nature of NEG_Peps, isolation of suppressor-of-phenotype clones on IPTG induced agar plates was readily possible. Suppressors in this study are defined as the genotypes evolved from the parental NEG_Pep1-6, such that they no longer show the prolonged growth phenotype and have not lost or mutated the NEG_Pep insert sequences or vectors. An important question that arises is, whether the suppressor clones can reveal the cellular targets of the deleterious NEG_Peps? NEG_Peps might cause the fitness defect by interacting with some cellular component or pathway by specific targeting and destabilizing the DNA, RNA or proteins. Mutations in these targets can be expected in the evolved suppressor genotypes that show a phenotypic rescue (i.e. no fitness defect). Three out of six parental NEG_Peps (i.e. NEG_Pep1, 2 and 5) gave rise to suppressor genotypes (three each Supp1, 2 and 3), all of which had mutations in the same positions (see chapter 3). Two of the mutational targets were in genes coding for outer membrane protein (*ydbA*) and tryptophanase (*tnaA*) whereas the third one was in the intergenic region upstream of the ribosomal RNA (*rrsH*). Hence, I speculate that the mutational targets in three suppressors can be the probable interacting partners of the corresponding NEG_Peps that they express, although this needs further experimental validation. Together the data show, that NEG_Peps may have potential targets in the host cellular machinery, which could cause the observed fitness defect. Fitness defect due to NEG_Peps is rescued in the suppressor genotypes, which may have acquired mutations in the gene coding the potential protein target, leading to loss of the initial interaction. Other mutations affecting the absolute peptide concentrations are also identified in some suppressors.

Another question arises at this point- how does the host cope with the deleterious effects of NEG_Peps? Organisms have an intrinsic ability to evolve, called evolvability (Sniegowski and Murphy, 2006) and they can adapt to environmental stressors by acquiring mutations. NEG_Peps and the corresponding suppressor re-sequencing data showed that there were two main categories of mutations in the genomic backgrounds: i) mutations affecting cellular pathways and ii) mutations affecting the plasmid (via copy number control or insert expression control). Several genes involved in the cell membrane physiology were the targets of suppressor mutations (six out of 12 suppressor genotypes; from the previously mentioned category ii.). For example, OmpN is a porin (NEG_Pep5 Supp1, chapter 3) that allows passive diffusion of small molecules outside the membrane (Fàbrega et al., 2012) and MntH is a metal ion transporter (NEG_Pep6 Supp3, chapter 3), both of which were targets in two suppressor genotypes.

6.1.3.2. NEG_Peps possibly cause concentration-dependent effects in host

Mutations affecting the plasmid are important because they directly control the levels of the NEG_Peps inside the cells. Among the re-sequenced suppressor genotypes, 11 out of 18 suppressor genotypes in the DH10B background (see chapter 4) had mutations in either *pcnB* (Poly A polymerase- PAP I) or *lacI* (lac repressor). Two strategies were seen in the suppressors with these mutations: i) mutations in the sugar binding region of the lactose repressor (*lacI*) gene (Wilson et al., 2007) and ii) mutations in the poly(A) polymerase gene which decreases the plasmid copy number (Lopilato et al., 1986). All mutations in the lac repressor were single nucleotide changes leading to non-synonymous mutations in the sugar binding cleft of the protein (Suckow et al., 1996). All mutations in LacI were Is type mutations, i.e. substitutions, which render the protein unresponsive to IPTG (Suckow et al., 1996). If the repression cannot be relieved, expression of NEG_Peps will not start and the cell would not accumulate the deleterious peptides. This is a direct control of the amount of NEG_Peps produced in the cells. Nevertheless, the genomic copy of *lacI* was apparently still active and would have thus allowed for low amounts of NEG_Peps to be produced, which was evident from presence of peptides in the western blots. Another mechanism which can reduce the absolute concentration of NEG_Peps inside the cells is an indirect control by PAP I. PAP I protein, when non-functional reduces the ColE1 origin plasmid copy number (see chapter 4). The mutations in the PAP I were mobile genetic element mediated insertions (IS elements), except in one case where it was a single nucleotide deletion. IS elements were found to be disrupting the *pcnB* gene along with target site duplication of some nucleotides. Disruption of the PAP I protein causes reduction in pFLAG (contains ampicillin cassette) vector containing the NEG_Peps. I showed this by testing the ampicillin susceptibility of the *pcnB* mutants, where all of them were more susceptible to increasing concentration of the antibiotic, although to different degrees. Since suppressor genotypes showed mutations that controlled the final accumulation of NEG_Peps inside the host, an immediate question arises whether NEG_Peps are also harmful at lower concentrations? NEG_Peps expressed from a low copy vector indeed did not show the deleterious phenotype. Together the results show that the NEG_Peps are tolerated at lower concentrations but are deleterious at higher concentrations. This hints towards a possible concentration dependent toxicity mechanism, which has been previously studied for dosage sensitive proteins of yeast (Bolognesi et al., 2016). It has been shown that above a critical over-expression threshold of certain proteins, a liquid-liquid phase separation occurs which causes cellular toxicity (Bolognesi et al., 2016). The NEG_Peps could also have a similar effect on their host cells where they are toxic at high concentrations leading to formation of foci that have higher concentration of the peptides possibly causing lethal liquid-liquid phase separation.

6.2. Future directions

The findings of this thesis have provided a proof of concept to the processes of *de novo* gene evolution. Based on this, it is possible to conclude that random nucleotides can serve as the raw materials for new gene functions and adaptations. Although a proof of concept was shown here, deeper understanding of the random peptide interaction and localization with the cellular machinery is still necessary. Insights into the interaction partners of each random peptide would be the immediate next step, which will advance the understanding of host proteins that are involved. In general, tagging the peptides with a fluorescent marker and following the localization and expression of different peptides will give more information about the function of these peptides before and after evolution of suppressors. It was shown in this study that the NEG_Peps affect the host by potentially affecting the membrane physiology. The response of each random sequence could potentially change under different conditions, which can be exploited to study conditional emergence of various beneficial sequences.

The beneficial effects conferred by random peptides could also be somewhat relevant in case of a competition scenario i.e. through a mechanism similar to clonal interference, where a beneficial variant can arise in an asexual population and spread (Fogle et al., 2008). This can be tested by competing the beneficial candidate against a mixture of neutral candidates and monitor the relative fitness. An important question arises of whether the beneficial random peptides provide a promiscuous advantage to the host. Testing the effects of beneficial random peptides by challenging the host in various stressful environments would shed light on the nature of the peptides. They could be investigated further by experimental evolution, wherein fluctuating environment (mimicking the wild) can be used each day and adaptation of host in the presence of beneficial peptides can be tested. This way one can observe if random sequences become an essential part of the host after selection, just like a new gene gaining a new function.

6.3. Concluding remarks

The insights gained from the empirical findings of this thesis suggests that random peptides can be functional in *E. coli*. This study has shown for the first time that individual random peptides can have at least a conditional fitness advantage in isolation. This study has also shown that random peptides can have severe fitness defects in *E. coli*, highlighting possible cellular interactions partners of the peptides. Furthermore, bacteria find diverse ways to cope with the deleterious peptides by evolving suppressor genotypes. Studying functionality of random sequences is not only vital to the process of new gene evolution, but it provides several potential uses in the therapeutic approaches. Specifically, if the deleterious random peptides are tested for their inhibitory properties they can provide us with new treatment strategies with the ever-growing problem of antibiotic resistance. Since the deleterious random peptides were found to be interacting with the proteins that govern the cell membrane physiology, it can be speculated that they may have similar implications on other host bacteria.

List Of Figures

List Of Tables

Bibliography

Alekshun MN, Levy SB (1997) Regulation of chromosomally mediated multiple antibiotic resistance: the mar regulon. Antimicrob Agents Chemother **41:** 2067-2075

Angyan AF, Perczel A, Gaspari Z (2012) Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: is aggregation the main bottleneck? FEBS Lett **586:** 2468-2472

Atsumi S, Wu TY, Machado IM, Huang WC, Chen PY, Pellegrini M, Liao JC (2010) Evolution, genomic analysis, and reconstruction of isobutanol tolerance in Escherichia coli. Mol Syst Biol **6:** 449

Bao Z, Clancy MA, Carvalho RF, Elliott K, Folta KM (2017) Identification of novel growth regulators in plant populations expressing random peptides. Plant physiology **175:** 619-627

Bartel DP, Szostak JW (1993) Isolation of new ribozymes from a large pool of random sequences [see comment]. Science **261:** 1411-1418

Begun DJ, Lindfors HA, Thompson ME, Holloway AK (2006) Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags. Genetics **172:** 1675-1681

Bolognesi B, Lehner B (2018) Reaching the limit. Elife **7**

Bolognesi B, Lorenzo Gotor N, Dhar R, Cirillo D, Baldrighi M, Tartaglia GG, Lehner B (2016) A Concentration-Dependent Liquid Phase Separation Can Cause Toxicity upon Increased Protein Expression. Cell Rep **16:** 222-231

Bradshaw AD (1965) Evolutionary significance of phenotypic plasticity in plants. Advances in genetics **13:** 115-155

Bulmer M (1990) The effect of context on synonymous codon usage in genes with low codon usage bias. Nucleic Acids Res **18:** 2869-2873

Cai J, Zhao R, Jiang H, Wang W (2008) De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. Genetics **179:** 487-496

Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, Brar GA, Weissman JS, Regev A, Thierry-Mieg N, Cusick ME, Vidal M (2012) Proto-genes and de novo gene birth. Nature **487:** 370-374

Chant EL, Summers DK (2007) Indole signalling contributes to the stable maintenance of Escherichia coli multicopy plasmids. Mol Microbiol **63:** 35-43

Chiarabelli C, Vrijbloed JW, Thomas RM, Luisi PL (2006) Investigation of de novo totally random biosequences, Part I: A general method for in vitro selection of folded domains from a random polypeptide library displayed on phage. Chemistry & biodiversity **3:** 827-839

Chothia C (1992) One thousand families for the molecular biologist. Nature **357:** 543-544

Christensen SK, Gerdes K (2003) RelE toxins from bacteria and Archaea cleave mRNAs on translating ribosomes, which are rescued by tmRNA. Mol Microbiol **48:** 1389-1400

Christensen SK, Mikkelsen M, Pedersen K, Gerdes K (2001) RelE, a global inhibitor of translation, is activated during nutritional stress. Proc Natl Acad Sci U S A **98:** 14328-14333

de Nadal E, Ammerer G, Posas F (2011) Controlling gene expression in response to stress. Nat Rev Genet **12:** 833-845

Deatherage DE, Barrick JE (2014) Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods Mol Biol **1151:** 165-188

Delaye L, Deluna A, Lazcano A, Becerra A (2008) The origin of a novel gene through overprinting in Escherichia coli. BMC Evol Biol **8:** 31

Delaye L, DeLuna A, Lazcano A, Becerra A (2008) The origin of a novel gene through overprinting in Escherichia coli. BMC Evolutionary Biology **8:** 31

Ding Y, Zhou Q, Wang W (2012) Origins of New Genes and Evolution of Their Novel Functions. Annual Review of Ecology, Evolution, and Systematics **43:** 345-363

Dobson CM (2003) Protein folding and misfolding. Nature **426:** 884-890

Domazet-Loso T, Brajkovic J, Tautz D (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet **23:** 533-539

Donnelly AE, Murphy GS, Digianantonio KM, Hecht MH (2018) A de novo enzyme catalyzes a life-sustaining reaction in Escherichia coli. Nat Chem Biol **14:** 253-255

Dornenburg JE, DeVita AM, Palumbo MJ, Wade JT (2010) Widespread antisense transcription in Escherichia coli. MBio **1:** e00024-00010

Dujon B (1996) The yeast genome project: what did we learn? Trends in Genetics **12:** 263-270

Ellington AD, Szostak JW (1990) In vitro selection of RNA molecules that bind specific ligands. nature **346:** 818

Engl C, Jovanovic G, Lloyd LJ, Murray H, Spitaler M, Ying L, Errington J, Buck M (2009) In vivo localizations of membrane stress controllers PspA and PspG in Escherichia coli. Molecular microbiology **73:** 382-396

Ermolaeva MD (2001) Synonymous codon usage in bacteria. Current issues in molecular biology **3:** 91-97

Fàbrega A, Rosner JL, Martin RG, Solé M, Vila J (2012) SoxS-dependent coregulation of ompN and ydbK in a multidrug-resistant Escherichia coli strain. FEMS Microbiol Lett **332:** 61-67

Famulok M, Szostak J (1993) Selection of Functional RNA and DNA Molecules from Randomized Sequences. *In* Nucleic Acids and Molecular Biology. Springer, pp 271-284

Felippes FF, Schneeberger K, Dezulian T, Huson DH, Weigel D (2008) Evolution of Arabidopsis thaliana microRNAs from random sequences. RNA **14:** 2455-2459

Fellner L, Bechtel N, Witting MA, Simon S, Schmitt-Kopplin P, Keim D, Scherer S, Neuhaus K (2014) Phenotype of htgA (mbiA), a recently evolved orphan gene of Escherichia coli and Shigella, completely overlapping in antisense to yaaW. FEMS microbiology letters **350:** 57-64

Fisher MA, McKinley KL, Bradley LH, Viola SR, Hecht MH (2011) De novo designed proteins from a library of artificial sequences function in Escherichia coli and enable cell growth. PLoS One **6:** e15364

Fogle CA, Nagle JL, Desai MM (2008) Clonal interference, multiple mutations and adaptation in large asexual populations. Genetics **180:** 2163-2173

Force A, Lynch M, Pickett FB, Amores A, Yan Y-l, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151:** 1531-1545

Ge SX, Jung D, Yao R (2019) ShinyGO: a graphical enrichment tool for animals and plants. Bioinformatics

Genest O, Hoskins JR, Camberg JL, Doyle SM, Wickner S (2011) Heat shock protein 90 from Escherichia coli collaborates with the DnaK chaperone system in client protein remodeling. Proc Natl Acad Sci U S A **108:** 8206-8211

Gerdes K, Bech FW, Jorgensen ST, Lobner-Olesen A, Rasmussen PB, Atlung T, Boe L, Karlstrom O, Molin S, von Meyenburg K (1986) Mechanism of postsegregational killing by the hok gene product of the parB system of plasmid R1 and its homology with the relF gene product of the E. coli relB operon. Embo j **5:** 2023-2029

Gerdes K, Poulsen LK, Thisted T, Nielsen AK, Martinussen J, Andreasen PH (1990) The hok killer gene family in gram-negative bacteria. New Biol **2:** 946-956

Gordon AJ, Burns PA, Fix DF, Yatagai F, Allen FL, Horsfall MJ, Halliday JA, Gray J, Bernelot-Moens C, Glickman BW (1988) Missense mutation in the lacI gene of Escherichia coli. Inferences on the structure of the repressor protein. J Mol Biol **200:** 239-251

Grassé P-P (2013) Evolution of living organisms: evidence for a new theory of transformation. Academic Press

Hajnsdorf E, Kaberdin VR (2018) RNA polyadenylation and its consequences in prokaryotes. Philos Trans R Soc Lond B Biol Sci **373**

Haldane J (1932) The causes of evolution. New York and London. *In*. Harper

Hall BG, Acar H, Nandipati A, Barlow M (2014) Growth rates made easy. Molecular biology and evolution **31:** 232-238

Haugel-Nielsen J, Hajnsdorf E, Regnier P (1996) The rpsO mRNA of Escherichia coli is polyadenylated at multiple sites resulting from endonucleolytic processing and exonucleolytic degradation. The EMBO journal **15:** 3144-3152

Heinen TJ, Staubach F, Haming D, Tautz D (2009) Emergence of a new gene from an intergenic region. Curr Biol **19:** 1527-1531

Higashi K, Ishigure H, Demizu R, Uemura T, Nishino K, Yamaguchi A, Kashiwagi K, Igarashi K (2008) Identification of a spermidine excretion protein complex (MdtJI) in Escherichia coli. J Bacteriol **190:** 872-878

Hochachka PW, Somero GN (2002) Biochemical adaptation: mechanism and process in physiological evolution. Oxford University Press

Hoegler KJ, Hecht MH (2016) A de novo protein confers copper resistance in Escherichia coli. Protein Sci **25:** 1249-1259

Hoegler KJ, Hecht MH (2018) Artificial gene amplification in Escherichia coli reveals numerous determinants for resistance to metal toxicity. Journal of molecular evolution **86:** 103-110

Horwitz M, Loeb LA (1986) Promoters selected from random DNA sequences. Proceedings of the National Academy of Sciences **83:** 7405-7409

Huvet M, Toni T, Sheng X, Thorne T, Jovanovic G, Engl C, Buck M, Pinney JW, Stumpf MP (2011) The evolution of the phage shock protein response system: interplay between protein function, genomic organization, and system function. Mol Biol Evol **28:** 1141-1155

Jacob F (1977) Evolution and tinkering. Science **196:** 1161-1166

Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res **20:** 1313-1326

Kaessmann H, Vinckenbosch N, Long M (2009) RNA-based gene duplication: mechanistic and evolutionary insights. Nature Reviews Genetics **10:** 19-31

Kaessmann H, Zöllner S, Nekrutenko A, Li W-H (2002) Signatures of domain shuffling in the human genome. Genome research **12:** 1642-1650

Kafri M, Metzl-Raz E, Jona G, Barkai N (2016) The Cost of Protein Production. Cell Rep **14:** 22-31

Keefe AD, Szostak JW (2001) Functional proteins from a random-sequence library. Nature **410:** 715-718

Keese PK, Gibbs A (1992) Origins of genes:" big bang" or continuous creation? Proceedings of the National Academy of Sciences **89:** 9489-9493

Khalturin K, Anton-Erxleben F, Sassmann S, Wittlieb J, Hemmrich G, Bosch TC (2008) A novel gene family controls species-specific morphological traits in Hydra. PLoS biology **6**

Kimura M, Ohta T (1974) On some principles governing molecular evolution. Proceedings of the National Academy of Sciences **71:** 2848-2852

Knopp M, Andersson DI (2018) No beneficial fitness effects of random peptides. Nat Ecol Evol **2:** 1046-1047

Knopp M, Gudmundsdottir JS, Nilsson T, Konig F, Warsi O, Rajer F, Adelroth P, Andersson DI (2019) De Novo Emergence of Peptides That Confer Antibiotic Resistance. mBio **10:** e00837-00819

Knowles DG, McLysaght A (2009) Recent de novo origin of human protein-coding genes. Genome Res **19:** 1752-1759

Kobayashi R, Suzuki T, Yoshida M (2007) Escherichia coli phage-shock protein A (PspA) binds to membrane phospholipids and repairs proton leakage of the damaged membranes. Molecular microbiology **66:** 100-109

Kornitzer D, Teff D, Altuvia S, Oppenheim AB (1991) Isolation, characterization, and sequence of an Escherichia coli heat shock gene, htpX. J Bacteriol **173:** 2944-2953

Lawrence JG, Hendrickson H (2003) Lateral gene transfer: when will adolescence end? Molecular microbiology **50:** 739-749

Lenski RE, Rose MR, Simpson SC, Tadler SC (1991) Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. The American Naturalist **138:** 1315-1341

Lerat E, Daubin V, Ochman H, Moran NA (2005) Evolutionary origins of genomic repertoires in bacteria. PLoS biology **3**

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ (2006) Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. Proc Natl Acad Sci U S A **103:** 9935-9939

Lewis M, Chang G, Horton NC, Kercher MA, Pace HC, Schumacher MA, Brennan RG, Lu P (1996) Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. Science **271:** 1247-1254

Li D, Dong Y, Jiang Y, Jiang H, Cai J, Wang W (2010) A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. Cell Res **20:** 408-420

Liberek K, Marszalek J, Ang D, Georgopoulos C, Zylicz M (1991) Escherichia coli DnaJ and GrpE heat shock proteins jointly stimulate ATPase activity of DnaK. Proceedings of the National Academy of Sciences **88:** 2874-2878

Liu J, Parkinson JS (1989) Genetics and sequence analysis of the pcnB locus, an Escherichia coli gene involved in plasmid copy number control. Journal of bacteriology **171:** 1254-1261

Long M, Betran E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. Nat Rev Genet **4:** 865-875

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 and General Genetics MGG **205:** 285-290

Luigi Luisi P (2003) Contingency and determinism. Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences **361:** 1141-1147

Maes A, Gracia C, Innocenti N, Zhang K, Aurell E, Hajnsdorf E (2016) Landscape of RNA polyadenylation in E. coli. Nucleic Acids Research **45:** 2746-2756

March J, Colloms M, Hart-Davis D, Oliver I, Masters M (1989) Cloing and characterization of an Escherichia coli gene, pcnB, affecting plasmid copy number. Molecular microbiology **3:** 903-910

Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. J Mol Biol **240:** 421-433

Martin RG, Jair K-W, Wolf R, Rosner JL (1996) Autoactivation of the marRAB multiple antibiotic resistance operon by the MarA transcriptional activator in Escherichia coli. Journal of bacteriology **178:** 2216-2223

Masai H, Arai K (1988) Initiation of lagging-strand synthesis for pBR322 plasmid DNA replication in vitro is dependent on primosomal protein i encoded by dnaT. J Biol Chem **263:** 15016-15023

Masai H, Arai K (1989) Escherichia coli dnaT gene function is required for pBR322 plasmid replication but not for R1 plasmid replication. J Bacteriol **171:** 2975-2980

Masters M, Colloms MD, Oliver IR, He L, Macnaughton EJ, Charters Y (1993) The pcnB gene of Escherichia coli, which is required for ColE1 copy number maintenance, is dispensable. Journal of Bacteriology **175:** 4405-4413

Miller JH, Schmeissner U (1979) Genetic studies of the lac repressor: X. Analysis of missense mutations in the lacI gene. Journal of molecular biology **131:** 223-248

Model P, Jovanovic G, Dworkin J (1997) The Escherichia coli phage-shock-protein (psp) operon. Mol Microbiol **24:** 255-261

Mohanty BK, Kushner SR (1999) Analysis of the function of Escherichia coli poly (A) polymerase I in RNA metabolism. Molecular microbiology **34:** 1094-1108

Mosca R, Pache RA, Aloy P (2012) The role of structural disorder in the rewiring of protein interactions through evolution. Molecular & Cellular Proteomics **11**

Muller HJ (1936) Bar duplication. Science **83:** 528-530

Murphy DN, McLysaght A (2012) De novo origin of protein-coding genes in murine rodents. PloS one **7**

Neme R, Amador C, Yildirim B, McConnell E, Tautz D (2017) Random sequences are an abundant source of bioactive RNAs or peptides. Nat Ecol Evol **1:** 0217

Neme R, Tautz D (2014) Evolution: dynamics of de novo gene emergence. Curr Biol **24:** R238-240

O'Hara EB, Chekanova JA, Ingle CA, Kushner ZR, Peters E, Kushner SR (1995) Polyadenylylation helps regulate mRNA decay in Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America **92:** 1807-1811

Ohno S (2013) Evolution by gene duplication. Springer Science & Business Media

Page R, Peti W (2016) Toxin-antitoxin systems in bacterial growth arrest and persistence. Nature chemical biology **12:** 208

Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden MT, Churcher CM, Bentley SD, Mungall KL (2003) Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. Nature genetics **35:** 32-40

Ptitsyn O (1985) Random sequences and protein folding. Journal of Molecular Structure: THEOCHEM **123:** 45-65

Ptitsyn O, Volkenstein M (1986) Protein structures and neutral theory of evolution. Journal of Biomolecular Structure and Dynamics **4:** 137-156

Regnier P, Hajnsdorf E (2009) Poly (A)-assisted RNA decay and modulators of RNA stability. Progress in molecular biology and translational science **85:** 137-185

Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD (2013) De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. PLoS Genet **9:** e1003860

Robertson DL, Joyce GF (1990) Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. Nature **344:** 467-468

Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Garner E, Guilliot S, Dunker AK (1998) Thousands of proteins likely to have long disordered regions. *In* Pac Symp Biocomput, Vol 3, pp 437-448

Sabath N, Wagner A, Karlin D (2012) Evolution of viral proteins originated de novo by overprinting. Mol Biol Evol **29:** 3767-3780

Saier MH (2000) A Functional-Phylogenetic Classification System for Transmembrane Solute Transporters. Microbiology and Molecular Biology Reviews **64:** 354-411

Sakoh M, Ito K, Akiyama Y (2005) Proteolytic activity of HtpX, a membrane-bound and stress-controlled protease from Escherichia coli. Journal of Biological Chemistry **280:** 33305-33310

Schlichting CD, Smith H (2002) Phenotypic plasticity: linking molecular mechanisms with evolutionary outcomes. Evolutionary Ecology **16:** 189-211

Schlotterer C (2015) Genes from scratch--the evolutionary fate of de novo genes. Trends Genet **31:** 215-219

Schmitz A, Schmeissner U, Miller JH (1976) Mutations affecting the quaternary structure of the lac repressor. journal of Biological Chemistry **251:** 3359-3366

Schroder H, Langer T, Hartl FU, Bukau B (1993) DnaK, DnaJ and GrpE form a cellular chaperone machinery capable of repairing heat-induced protein damage. EMBO J **12:** 4137-4144

Seelig B, Szostak JW (2007) Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. Nature **448:** 828-831

Sharp PM, Li W-H (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic acids research **15:** 1281-1295

Shaw D, Walker J, Northrop F, Barrell B, Godson G, Fiddes J (1978) Gene K, a new overlapping gene in bacteriophage G4. Nature **272:** 510-515

Siguier P, Filée J, Chandler M (2006) Insertion sequences in prokaryotic genomes. Current opinion in microbiology **9:** 526-531

Skowyra D, Georgopoulos C, Zylicz M (1990) The E. coli dnaK gene product, the hsp70 homolog, can reactivate heat-inactivated RNA polymerase in an ATP hydrolysis-dependent manner. Cell **62:** 939-944

Smyth GK (2005) Limma: linear models for microarray data. *In* Bioinformatics and computational biology solutions using R and Bioconductor. Springer, pp 397-420

Sniegowski PD, Murphy HA (2006) Evolvability. Current Biology **16:** R831-R834

Stepanov VG, Fox GE (2007) Stress-driven in vivo selection of a functional mini-gene from a randomized DNA library expressing combinatorial peptides in Escherichia coli. Mol Biol Evol **24:** 1480-1491

Suckow J, Markiewicz P, Kleina LG, Miller J, Kisters-Woike B, Müller-Hill B (1996) Genetic studies of the lac repressor XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. Journal of molecular biology **261:** 509-523

Tautz D (2014) The discovery of de novo gene evolution. Perspect Biol Med **57:** 149-161

Tautz D, Domazet-Loso T (2011) The evolutionary origin of orphan genes. Nat Rev Genet **12:** 692-702

Tomizawa J-i, Itoh T, Selzer G, Som T (1981) Inhibition of ColE1 RNA primer formation by a plasmid-specified small RNA. Proceedings of the National Academy of Sciences **78:** 1421-1425

Tomizawa J-i, Som T (1984) Control of cole 1 plasmid replication: enhancement of binding of RNA I to the primer transcript by the rom protein. Cell **38:** 871-878

Tretyachenko V, Vymětal J, Bednárová L, Kopecký V, Hofbauerová K, Jindrová H, Hubálek M, Souček R, Konvalinka J, Vondrášek J (2017) Random protein sequences can form defined secondary structures and are well-tolerated in vivo. Scientific reports **7:** 1-9

Uehara T, Suefuji K, Valbuena N, Meehan B, Donegan M, Park JT (2005) Recycling of the anhydro-N-acetylmuramic acid derived from cell wall murein involves a two-step conversion to N-acetylglucosamine-phosphate. J Bacteriol **187:** 3643-3649

Vandecraen J, Chandler M, Aertsen A, Van Houdt R (2017) The impact of insertion sequences on bacterial genome plasticity and adaptability. Crit Rev Microbiol **43:** 709-730

Vega NM, Allison KR, Khalil AS, Collins JJ (2012) Signaling-mediated bacterial persister formation. Nat Chem Biol **8:** 431-433

Verstraeten N, Knapen WJ, Kint CI, Liebens V, Van den Bergh B, Dewachter L, Michiels JE, Fu Q, David CC, Fierro AC (2015) Obg and membrane depolarization are part of a microbial bet-hedging strategy that leads to antibiotic tolerance. Molecular cell **59:** 9-21

Vulin C, Leimer N, Huemer M, Ackermann M, Zinkernagel AS (2018) Prolonged bacterial lag time results in small colony variants that represent a sub-population of persisters. Nat Commun **9:** 4074

Wade JT, Grainger DC (2014) Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Microbiol **12:** 647-653

Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. Nature **449:** 54-61

Weisman CM, Eddy SR (2017) Gene Evolution: Getting Something from Nothing. Curr Biol **27:** R661-R663

Wild J, Altman E, Yura T, Gross CA (1992) DnaK and DnaJ heat shock proteins participate in protein export in Escherichia coli. Genes Dev **6:** 1165-1172

Wild J, Rossmeissl P, Walter WA, Gross CA (1996) Involvement of the DnaK-DnaJ-GrpE chaperone team in protein secretion in Escherichia coli. J Bacteriol **178:** 3608-3613

Wilson BA, Foy SG, Neme R, Masel J (2017) Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth. Nat Ecol Evol **1:** 0146-0146

Wilson BA, Masel J (2011) Putatively noncoding transcripts show extensive association with ribosomes. Genome Biol Evol **3:** 1245-1252

Wilson CJ, Zhan H, Swint-Kruse L, Matthews KS (2007) The lactose repressor system: paradigms for regulation, allosteric behavior and protein folding. Cell Mol Life Sci **64:** 3-16

Wilson CJ, Zhan H, Swint-Kruse L, Matthews KS (2007) Ligand interactions with lactose repressor protein and the repressor-operator complex: The effects of ionization and oligomerization on binding. Biophysical chemistry **126:** 94-105

Wilson DS, Szostak JW (1999) In vitro selection of functional nucleic acids. Annual review of biochemistry **68:** 611-647

Wu D-D, Irwin DM, Zhang Y-P (2011) De novo origin of human protein-coding genes. PLoS genetics **7**

Xiao W, Liu H, Li Y, Li X, Xu C, Long M, Wang S (2009) A rice gene of de novo origin negatively regulates pathogen-induced defense response. PLoS One **4:** e4603

Xie C, Zhang YE, Chen J-Y, Liu C-J, Zhou W-Z, Li Y, Zhang M, Zhang R, Wei L, Li C-Y (2012) Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. PLoS genetics **8**

Xu F, Lin-Chao S, Cohen SN (1993) The Escherichia coli pcnB gene promotes adenylylation of antisense RNAI of ColE1-type plasmids in vivo and degradation of RNAI decay intermediates. Proceedings of the National Academy of Sciences **90:** 6756-6760

Yona AH, Alm EJ, Gore J (2018) Random sequences rapidly evolve into de novo promoters. Nat Commun **9:** 1530

Yu CH, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, Liu Y (2015) Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. Mol Cell **59:** 744-754

Zhai Y, Saier MH, Jr. (2002) The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes. Protein Sci **11:** 2196-2207

Ziemienowicz A, Skowyra D, Zeilstra-Ryalls J, Fayet O, Georgopoulos C, Zylicz M (1993) Both the Escherichia coli chaperone systems, GroEL/GroES and DnaK/DnaJ/GrpE, can reactivate heat-treated RNA polymerase. Different mechanisms for the same activity. J Biol Chem **268:** 25425-25431

Acknowledgements

It has been a great pleasure to be working with my supervisor Prof. Dr. Diethard Tautz. Your endless ideas and support made my PhD a successful and enjoyable journey. My PhD has been a rewarding ride only because of the constant help and encouragement I have received from countless folks throughout my time in Germany. Thank you Dr. Jenna Gallie for your guidance and persistent experimental help whenever needed. I owe a big thank you to Dr. Guy Reeves, who has provided numerous ideas and has asked the toughest questions about my results, which helped me build an effective story. Thanks Guy! I would like to thank Rafik Neme for his help and advice. I am thankful for my thesis committee members: Prof. Dr. Ruth Schmitz-Streit, Prof. Dr. Paul Rainey and Dr. Jenna Gallie, who helped me mould this thesis in the best possible way. I owe a big thanks to Prof. Dr. Rees Kassen, who gave me the temperature idea during our extensive conversation about science and science policies at the MPI Ploen. I would like to thank Ellen McConnel for helping me with the initial experiments. Thank you Kristian Ulrich for your support and ideas during the initial year of my PhD. I would also like to thank the post-docs of our Evolutionary Genetics department, each of who have helped me during my PhD. I am grateful to Prof. Dr. Hinrich Schulenburg, Dr. Camilo Barbosa and Dr. Kerstin Mehnert for ensuring my smooth transition into this PhD position. I would like to acknowledge the technicians at MPI (specially Elke Blohm-Sievers) who helped me with various experiments. I thank Dr. Sven Kuenzel for performing whole genome re-sequencing at the MPI. I would like to thank the administrative staff (especially Sabine, Angela, Marita, Katharina and Petra) at MPI, who helped me with the paperwork at the institute. I would also like to thank Derk Wachsmuth and the IT department for their support. Thank you Iben for your support in the library. I would like to thank Jenna, Guy, Anuradha, Juan, Suhrid, Maria and certainly Diethard, for providing with meticulous suggestions for improving this thesis.

Thank you Dr. Shashikant Acharya for providing me with a much needed kick-start into research after Masters. I owe Dr. Sutirth Dey and Dr. Shraddha Karve a debt of gratitude for driving my passion towards evolutionary biology. Thank you Population Biology Lab for the support during my IISER days!

All my friends have significantly contributed to what I have accomplished here. Anuradha, you have been an incredible friend and my PhD wouldn't have been this much fun if it weren't for you. Thank you for diligently coming up with improvised recipes each time for when I listed random items from my refrigerator! You have been a huge support. Johana Fajardo, you have been an important part of my PhD journey. Cheers to our endless discussions about *de novo* genes to life (with us, it always escalates quickly!). Thank you Alejandro, Loukas and Derek for being there. Thank you Maria and Juan for the motivation and company to all our sporty endeavours during these years. Cheers to our swimming trio and passionate late night full moon updates! Thank you Filipa, Gillian, Ezgi, Samer, Mayra, Malavi, Hyejin, Cecilia and Bilal for great conversations over coffees, beers, cocktails, etcetera—truly memorable times!

Without my family, the completion of this thesis would not have been possible. I fall short of words to describe the immense support my family has provided me during every step of my life as a PhD and beyond. I would like to thank my parents, Priyashree and Sanjay, my dear little sister, Kruttika and my grandfather, Vasant for always being there for me. My mother in particular, has been a great source of inspiration for me, thank you Aai. Last but not the least, I owe a big thanks to Suhrid—a remarkable scientist, friend and partner. Your unconditional love and support continue to give me the strength to endure tough times.

Declaration

I hereby declare that:

i. Apart from my supervisor's guidance the content and design of the paper is all my own work;

ii. This thesis has not been submitted either partially or wholly as part of a doctoral degree to another examining body and no material has been published or submitted for publication;

iii. The preparation of this thesis has been subjected to the Rules of Good Scientific Practice of the German Research Foundation;

iv. No academic degree has ever been withdrawn prior to this thesis.

Ploen, 23rd June 2020

Devika Bhave