

Deep mutational scanning of an alanine
racemase from *Streptococcus*
thermophilus



UCLouvain

Jason Baby Chirakadavil

Louvain Institute of Biomolecular Science and Technology
Université catholique de Louvain

A thesis submitted for the degree of
PhD in Science
2021

Dedication

This thesis is dedicated to my late grandpa **Samuel Varghese**.

Appachan, you will always be in our hearts and minds.

Members of the jury

President of the jury : **Prof. François Chaumont**

Université catholique de Louvain – Louvain Institute of Biomolecular
Science and Technology

Reviewer : **Prof. Wayne Patrick**

Victoria University of Wellington – School of Biological Sciences

Reviewer : **Prof. Clement Nizak**

Sorbonne Université – Laboratoire Jean Perrin

Reviewer : **Prof. Pierre Morsomme**

Université catholique de Louvain – Louvain Institute of Biomolecular
Science and Technology

Co-supervisor : **Prof. Pascal Hols**

Université catholique de Louvain – Louvain Institute of Biomolecular
Science and Technology

Supervisor : **Prof. Patrice Soumilion**

Université catholique de Louvain – Louvain Institute of Biomolecular
Science and Technology

Summary

It is of fundamental interest in evolutionary biology to link the effect of mutations in a protein to the fitness of an organism. One of the best methods to investigate this effect is to perform a deep mutational scan (DMS). This involves creating a library of protein mutants which are subject to selection in diverse conditions and consequently high throughput sequencing is employed to monitor the fate of mutants. In this study, in the first section we introduce a novel method of creating large chromosomally encoded libraries (upto 10^9 variants) by a two-step scarless gene replacement using a *cat-oroP* cassette in *Streptococcus thermophilus*. We use this method to create a library of 10^8 variants of alanine racemase (Alr) cloned at its native locus in the chromosome. Then in the second section, the Alr library was grown in two conditions, in presence and absence of selection pressure for function. Growth under functional selection, created a map of residues essential to alanine racemase activity expressed under native conditions. We observe that 44% of substitutions were deleterious for activity. In absence of selection pressure for function, we observe that the majority of substitutions (67%) had neutral effects on fitness. In both conditions, certain substitutions were selected which may be involved in restricting entry to the active site. In the third section, we engineered a *S. thermophilus* strain by inserting a SARS-CoV-2 protease recognition sequence in the Alr protein. This mutant displayed conditional auxotrophy on expressing the protease. This strain was used to screen libraries of peptidic inhibitors to identify SARS-CoV-2 protease inhibitors.

Acknowledgements

This thesis has been quite an amazing journey and now close to the destination, I would like to thank the people without whom this wouldn't have been possible.

I would like to first say a big thank you to my supervisor **Prof. Patrice**, without whom none of this would have been possible. Thank you for your amazing curiosity about science, thank you for being a great friend, mentor, and supervisor. You have made my time in the lab absolutely enjoyable. Many thanks to my co-supervisor **Prof. Pascal**, from the first day I knew I could walk into your office and ask you anything and you would know it right away. Thank you for all the help. I would also like to extend a huge thank you to **Prof. Chaumont, Prof. Wayne, Prof. Nizak and Prof. Morsomme** for reading my manuscript and providing me with incredible suggestions.

The BGM has been the most welcoming lab I have ever been at. I have nothing but the highest praise for my best friend, science guru and random blabber-mate, beaver-watching, cycling, hiking, camping, house and office mate of four years **Sebastian Worms**. Seb is an amazing person in way too many ways. I will miss you a lot, buddy. I thank you, my PhD buddy, **Gol Dorrazehi**, it was quite an amazing journey we started together as Marie-Curie ESRs. Thank you **Steve Alarcia** for being such a great friend and python guru! I can't thank you enough for helping me with all the data analysis. I hope someday I get to share an office with you guys again.

A special group which deserves their own section, because of the big part they played in my life. **Stefano Di Giacomo** '*my dear papi*', **Serena Garguilo** and **Laura Ledesma**, I am happy to have met you guys and made such close friends. I hope that we continue to be friends for the coming years. Thank you to **Adrian Knoops** and for introducing me to my second-love, the wonderful world of outdoor rock climbing and to **Benoît Desguin** for introducing me to skiing, for being such a great friend, I loved all our trips. I hope I can continue to join all the trips from Lyon. Thank you to **Zia Tariq** and **Tessa de Bruin** for the discussions and long conversations over lunch and coffee. Thank you to **Johann Mignolet** for all the advice regarding with thesis organization and lab work and for organizing the amazing poker nights. A big thank you to **Salwa Ben Abbou** for her help during long weeks preparing the alanine racemase library. Finally a big thank you to **Brigitte Delplace** and **Damien Evrard** for keeping the whole lab running and all our projects afloat, I have no idea how anything would/could work without you guys, also a collective sorry from all of us for the endless little emergencies that you both have to deal with. I thank **Bernard Hallet** for always having had interesting questions during team presentations, and always being there to answer all my doubts. Finally a big thank you to my friends, the current and ex- members of **BGM**, **Julien**, **Perrine**, **Estelle**, **Shahnawaz**, **Frédéric**, **both Ophélies**, and **all the members of the amazing BGM team**, you guys made daily life in the lab so enjoyable.

I'd like to thank all of you who are closest to me. A big thank you to **Francesca** my closest friend and housemate. Thank you for the long walks in the evenings, for chasing the maison carette ice cream trucks, and for being my horse riding and cooking partner and thank you for being a constant source of motivation, love and support during my writing. Thank you to **Pierre**, **Claire** and **Francesca** for building such an amazing

colocation with me, for all the amazing memories, for building the pond and the potager, thank you **Claire** for rekindling in me, the amazing love for plants and wildlife, and thank you **Pierre** for introducing me the incredible Brewery de l'Ermitage, and teaching about the world of Kant, Descartes, Epictetus... Thank you, **Pascale Hachez** for almost being my Belgian *mom* and **Valentine** for making me feel at home from the very beginning.

Finally India, to my pre-UCL days. Herein lies my most basic source of my energy. Thank you **Mom** and **Dad**, for always being there for me. Thank you your calls every single day for 4 years always making sure I was good, thank you for all the sacrifices you made so that me and James could be where we are today. Thank you for your love and support, my amazing brother **James**, you have always been a close friend and my hero and you've practically been a second *dad* to me. Thank you **Steffi** for all your love and care during all these years. Additionally, thank you **Pinky mami**, **Suja ammama**, **Jossy achacha**, **Suku mama**, **Joel**, **Rachel**, **Samson**, **Johnson** for always bring there for me.

I end with my BFFs, my *purane aur hamesha ke dost*. Thank you **Dushyant** and **Atharva**, although continents separate us we've just grown closer over the years, 'आगर तुम दोनों हो तोह और क्या यह दिल चाहेगा'. Thank you to my dear **Sravya**. *Yede*, thank you for always being there, no matter what hour it is, you've been there for me throughout the years. You are my Rock of Gibraltar. Finally, thank you **Dhwani**, you helped me look for this position, hopefully we can have *cheese momos* someday at Appetite again.

Thank you **Ajinkya**, **Chintan**, **Smruti**, **Swetha**, **Ashwin**, **Chaitra**.

In these four years, I have evolved, alongside my bacteria. I walked out of Brussels airport four years back, naïve and unprepared. Each one of the people mentioned above moulded and supported me, I would never have reached here without you guy

Funding acknowledgements

I gratefully acknowledge the funding and support received towards my PhD as an **Marie Curie - Early Stage Researcher (ESR)**.

This project has received funding from the **European Union's Horizon 2020** research and innovation programme under the **Marie Skłodowska-Curie grant agreement No 722610 (ES-Cat)**.

This material reflects only the author's views and the Union is not liable for any use that may be made of the information contained therein.



List of acronyms

• CN	Contact number
• WCN	Weighted contact number
• RSA	Relative solvent accessibility
• NGS	Next generation sequencing
• DMS	Deep mutational scanning
• GPCR	G protein coupled receptor
• Fab	Fragment antigen binding
• YAP65	Yes-associated protein-65
• PKA	Protein kinase A
• PSD95	Postsynaptic density protein-95
• Hsp90	Heat-shock protein 90
• Mat α 2	Mating type protein- α 2
• Pab1	Poly(A)binding protein-1
• CDR	Complementarity determining region
• Bgl-3	Beta-glucosidase
• β 2AR	β 2-adrenergic receptor
• GRAS	Generally recognized as safe
• EFSA	European Food Safety Authority
• QPS	Qualified Presumption of Safety
• ROS	Reactive oxygen species

• XIP	ComX-inducing peptide
• PLP	Pyridoxal-5-phosphate
• Alr	Alanine racemase
• PDB	Protein data bank
• PCR	Polymerase chain reaction
• DNA	Deoxyribonucleic acid
• 5-FOA	5-Fluoroorotic acid
• dTMP	Deoxythymidine monophosphate
• Cat	Chloramphenicol acetyl transferase
• OD	Optical density
• RLU	Relative light unit
• UMP	Uridine monophosphate
• CFU	Colony forming unit
• SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
• SICLOPPS	Split intein-mediated circular ligation of peptide and proteins
• tRNA	transfer RNA
• ATCC	American Type Culture Collection
• DMSO	Dimethyl sulfoxide
• CDM	Chemically defined medium

Table of contents

1.	Introduction	1
1.1.	Molecular evolution	2
1.2.	Deep mutational Scanning	8
1.3.	<i>Streptococcus thermophilus</i> as model organism	17
1.4.	Alanine racemase as a model enzyme	21
1.4.1.1.	Structure of alanine racemase	25
1.4.1.2.	Reaction mechanism	34
2.	Aim of the project and strategy	37
3.	Results	41
3.1.	<i>Streptococcus thermophilus</i> as a suitable host for constructing large scarless variant libraries cloned and expressed from the chromosome	41
3.1.1.	Abstract	42
3.1.2.	Introduction	43
3.1.3.	Results	46
3.1.4.	Discussion	61

3.1.5.	Future Work	62
3.2.	Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in <i>Streptococcus thermophilus</i>	65
3.2.1.	Abstract	66
3.2.2.	Introduction	67
3.2.3.	Results 1	73
3.2.4.	Discussion 1	122
3.2.5.	Results 2	129
3.2.6.	Discussion 2	142
3.3.	Engineering <i>Streptococcus thermophilus</i> as a reporter strain for selecting peptidic inhibitors of SARS-CoV-2 Mpro and PLpro proteases	147
3.3.1.	Abstract	148
3.3.2.	Introduction	148
3.3.3.	Results	152
3.3.4.	Discussion	167
3.3.5.	Future Work	172
4.	Materials and Methods	

Table of contents

4.1.	Building Scarless Gene Libraries in the Chromosome of Bacteria	176
4.2.	Materials and Methods for Chapter 3.1	195
4.3.	Materials and Methods for Chapter 3.2	198
4.4.	Materials and Methods for Chapter 3.3	203
4.5.	Materials and Methods: Appendices	207
5.	Bibliography	221

1.

Introduction

Understanding the effects of mutations on the function of a protein is one of the fundamental areas of inquiry in molecular evolution. Existing biophysical and evolutionary models cannot explain all these effects and there is a need for experimental approaches to observe major and more subtle impacts of mutations in a high throughput way and under conditions that simulate natural evolution as well as possible. This project focusses on the study of evolution of a library of alanine racemase variants cloned in its natural chromosomal locus as a means to explore the fundamental selection pressures at the molecular level.

This introduction is organised into four sections. We will start by presenting (1.1) some of the unanswered questions in molecular evolution. In the second section (1.2), we will describe the state-of-the-art of high throughput mutagenesis techniques collectively termed as '*deep mutational scanning*'. The third section (1.3) is an overview of benefits of the model organism used in this study, *Streptococcus thermophilus*, while the last one (1.4) is presenting the alanine racemase that was chosen as the model enzyme for genotype-phenotype mapping.

1.1. Molecular evolution

Evolution of proteins depends on an interplay of biophysical constraints (like solvent accessibility, thermal stability etc.) with functional constraints (related to catalytic activity, interactions with other cellular entities). Additionally, cells are constantly facing changing environmental conditions and new functional constraints emerge under evolutionary pressures such as antibiotic resistance, exposure to non-natural compounds (Seffernick et al. 2001), etc. Can the fate of every single amino acid substitutions be explained by already 'known' constraints on rate of evolution? The answer is clearly no and one of the interesting questions we still need to explore is what are the constraints explaining evolutionary rate variation in proteins. **Table 1** lists the effects of known constraints on site-specific rate variation in proteins.

Introduction

However, we know that current prediction models based on these existing biophysical constraints can explain only upto 60% of rate variation (Dean et al. 2002; Echave, Spielman, and Wilke 2016).

Table 1. Effects of known constraints on rate variation in proteins.

Type of constrain	Effect on rate variation	Reference
Structural constraints :		
CN/WCN	Reduces with increasing CN/WCN	(Franzosa and Xia 2009; Marcos and Echave 2015; Shahmoradi et al. 2014; Shih et al. 2012; Yeh et al. 2014)
RSA	Increases with increasing RSA	(Bustamante, Townsend, and Hartl 2000; Goldman, Thorne, and Jones 1998; Noël and Sermet-Gaudelus 2020; Ramsey et al. 2011)
Structural flexibility	Increases with increasing flexibility	(T. T. Huang et al. 2014; Nevin Gerek, Kumar, and Banu Ozkan 2013; Shahmoradi et al. 2014)

Structural disorder	Increases in disordered regions	(Brown et al. 2002; Tóth-Petróczy and Tawfik 2011)
Functional constraints :		
Protein interaction interfaces	Reduced in interface regions	(Franzosa and Xia 2009; P. M. Kim et al. 2006; Mintseris and Weng 2005)
Catalytic sites	Reduced at and near the site	(Dean et al. 2002; Jack et al. 2016)
Gene expression level	Decreases with increasing expression level	(Drummond et al. 2005)

CN, Contact number is the number of other residues in contact with a given residue. WCN, weighted contact number. RSA, Relative solvent accessibility. Modified from (Echave, Spielman, and Wilke 2016).

Introduction

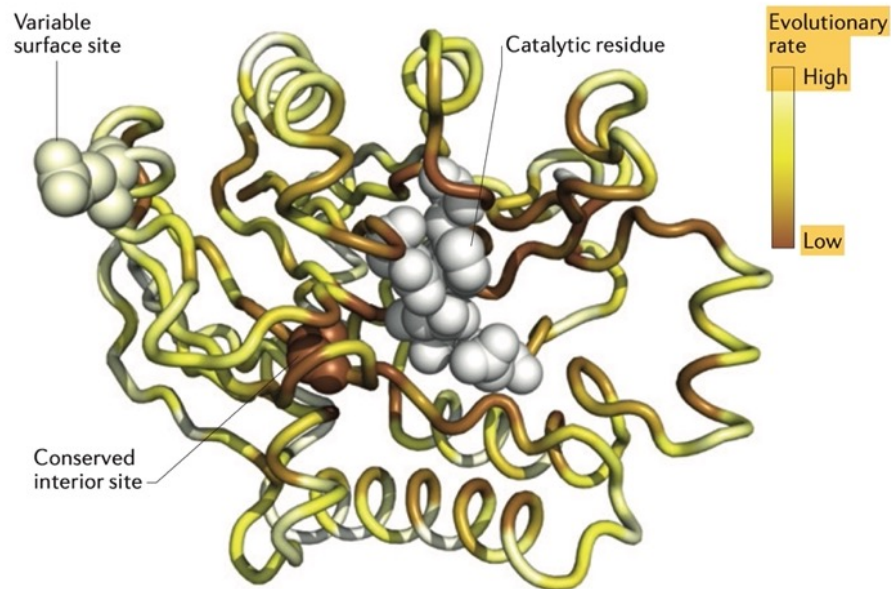


Figure 1. Structural and functional constraints shape evolutionary divergence at sites. Substitution rates for *E. coli* Exonuclease III were mapped onto the structure. It can be seen clearly that substitution rates are low in the interior of the protein owing to structural constraints (RSA, etc) and higher on exposed residues at the surface of the protein. Additionally, residues at and near the catalytic site also evolve slowly indicating functional constraints (catalytic activity, protein DNA interface). Figure taken from (Echave, Spielman, and Wilke 2016) where the evolutionary rate data is from (Marcos and Echave 2015).

Can natural sequence conservation identify the effect of mutations?

According to Boucher et al., for a protein for which the function is conserved, selection would predominantly act to maintain that

function. Over the long evolutionary timeline, mutations with very small deleterious effects (e.g., 0.1%) and those with strong deleterious effects (e.g., 100%) would both be completely purged from large natural populations. Conserved residues in an alignment of homologous protein sequences indicate purifying selection, but it cannot delineate the strength of the fitness defect. On the other hand, the strength of fitness effects can be measured experimentally, although usual experimental methods are inherently time consuming. Here, high throughput assays can play a complementary role, adding to information obtained from sequence alignment and biochemical assays. For example, in the systematic analysis of ubiquitin mutations by Roscoe et al., sequence conservation showed the entire ubiquitin sequence very strictly conserved with only 4 sites showing variation, but experimental results painted a slightly different picture. Although the binding interfaces demonstrated high conservation, most surface positions were very tolerant in contrast with sequence alignments data (Roscoe et al. 2013). Two non-exclusive and probable reasons may explain this difference between natural sequence conservation and experimental fitness measurements : (1) fitness effects subject to selection in large natural populations are beyond the resolution power of current experiments and/or (2) experimental conditions are not as stringent as evolution in the natural context which is subjected to more nuanced selection pressures in diverse and changing environmental conditions. Supporting this second

Introduction

hypothesis, Jiang et. al. have demonstrated that mimicking natural expression conditions for Hsp90 create a better nuanced picture of selection when compared to experiments with artificially high expression levels (Boucher et al. 2014; Jiang et al. 2013).

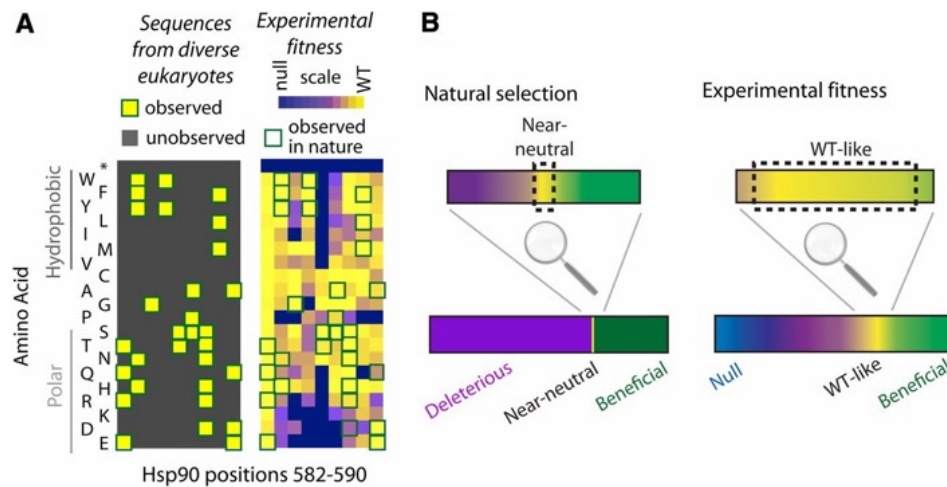


Figure 2. (A) Heatmap of fitness effect of sequence conservation between sequences from many eukaryotes compared to experimentally observed fitness values for the same region. Hsp90 function is strongly conserved in eukaryotes (Picard et al. 1990). (B) The observation of nearly neutral mutations is cryptic in natural sequence alignments due to purifying selection. Slightly deleterious (e.g., 1% defect) and strongly deleterious mutations (e.g., 100 defect) should both be completely purged from population. The near neutral window of fitness effect in nature is theoretically proportional to the inverse of the population size (Ohta 1973). Experimental approaches can reveal the full range of fitness effects

but often fail to recognise the window of near neutral mutations, due to limited stringency in the lab conditions. Taken from (Boucher et al. 2014)

Some fundamental questions in molecular evolution are still unanswered with current experimental systems,

- How much of the unexplained variation in site rate can be explained due to errors, artefacts or biases in estimation ?
- Are there hidden selection pressures we don't know about, and can these explain some of the rate variation?
- Is there a threshold of protein stability above which all stability-affecting mutations are neutral?
- How does the landscape of neutrally drifted genotypes look like?
- How is the freedom to evolve when a protein is under constraint for native function?

The combination of high throughput protein fitness analysis with carefully designed experimental conditions mimicking diverse natural constraints and selection could answer some of these questions. *In vivo* deep mutational scanning can be performed in diverse conditions. The next section will give an overview of reported strategies before discussing the benefits of deep mutational scanning for exploring fundamental evolutionary questions.

1.2. Deep mutational scanning

In the last decade, a set of sequencing techniques collectively known as next generation sequencing (NGS) have been developed,

Introduction

allowing sequencing millions of DNA sequences in parallel. One of the applications of NGS sequencing is the measurement of DNA abundance as a quantitative output for measuring biological phenomenon like gene expression, protein-nucleic acid interactions, chromatin accessibility, RNA modifications (Araya and Fowler 2011; Mikkelsen et al. 2007; Nagalakshmi et al. 2008; S. B. Ng et al. 2009, 2010; Sultan et al. 2008). Recently, functional selection has been coupled with NGS to measure the effect of mutations on protein activity. In these methods, a library of protein variants is first created such that the phenotype under study is directly associated to the genotype. The library is then subjected to selection (enzyme activity, binding to ligand etc.) and the library is sequenced by NGS before and after selection. Finally, the enrichment or purge of protein variants is estimated by measuring the abundance of DNA (Figure 3). This method of estimating protein activity has been reviewed broadly in Araya & Fowler, 2011 and Fowler & Fields, 2014, and has been termed as 'deep mutational scanning' since 2014.

Many previous studies over the years have also used the same techniques to study many different properties of various proteins. The list presented in **Table 2** is not a comprehensive list, but it highlights the diversity of proteins and selection systems employed in deep mutational scanning experiments.

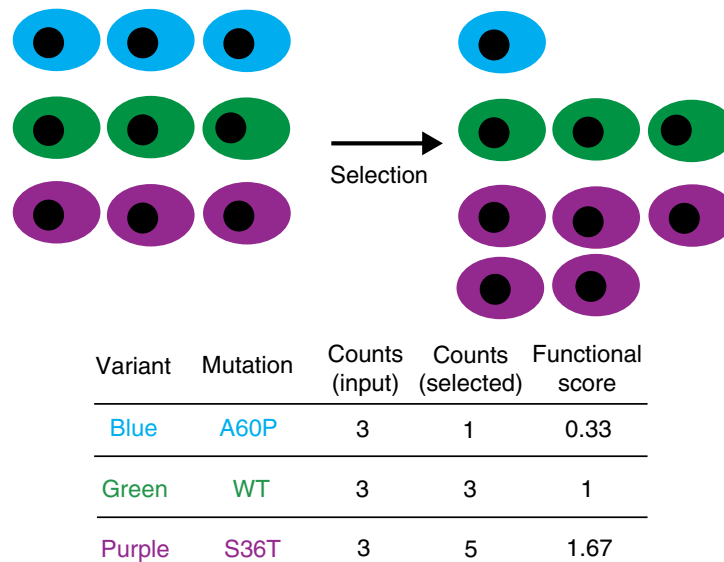


Figure 3. Measurement of protein activity by deep mutational scanning.

A library of protein variants is created such the variant phenotype is linked to the genotype. This library is sequenced by NGS before and after selection. Variants with the selected phenotype will increase in abundance after selection, while other variants are purged and will decrease in abundance. This data is analysed to generate a functional score for each variant. Blue, green and purple ovals represent the different variants. Taken from (Fowler and Fields 2014).

Table 2. Deep mutational scanning in recent literature

Scanned protein	Model system	Selection	Reference
-----------------	--------------	-----------	-----------

Introduction

Fab antibody fragment	Ribosome display	Ligand binding	(Fujino et al. 2012)
YAP65 WW domain	T7 bacteriophage	Ligand binding	(Starita et al. 2013)
E4B ubiquitin ligase	T7 bacteriophage	Ubiquitination activity	(Araya et al. 2012)
PKA regulatory subunit	T7 bacteriophage	Ligand binding	(Gold et al. 2006)
Synthetic PDZ domain	M13 bacteriophage	Ligand binding	(Ernst et al. 2010)
CcdB	<i>Escherichia coli</i>	Toxin activity	(Adkar et al. 2012)
PSD95 PDZ domain	<i>E. coli</i>	Ligand binding	(McLaughlin et al. 2012)
GPCR	<i>E. coli</i>	Ligand binding	(Schlinkmann et al. 2012)
Designed influenza inhibitor	<i>Saccharomyces cerevisiae</i> surface display	Ligand binding	(Whitehead et al. 2012)
Designed lysozyme inhibitor	<i>S. cerevisiae</i> surface display	Ligand binding	(Procko et al. 2013)
Designed digoxigenin binder	<i>S. cerevisiae</i> surface display	Small molecule binding	(Tinberg et al. 2013)
IgG1 CH3 domain	<i>S. cerevisiae</i> surface display	Ligand binding after thermal stress	(Traxlmayr et al. 2012)
Hsp90	<i>S. cerevisiae</i> complementation	Growth rate	(Ryan T Hietpas, Jensen, and Bolon 2011;

			Jiang et al. 2013)
Mata2 degron	<i>S. cerevisiae</i> fusion protein	Growth rate	(I. Kim et al. 2013)
Ubiquitin47	<i>S. cerevisiae</i> complementation	Growth rate	(Roscoe et al. 2013)
Pab1	<i>S. cerevisiae</i> complementation	Growth rate	(Melamed et al. 2013)
Neuraminidase	Mammalian cell	Oseltamivir resistance	(N. C. Wu et al. 2013)
IgG CDRs	Mammalian cell	Growth rate	(Forsyth et al. 2013)
B-Raf	Mammalian cell	Vemurafenib resistance	(Wagenaar et al. 2014)
HIV Envelope protein	HIV provirus	Viral infectivity	(Haddox et al. 2018)
Zika Envelope protein	Zika provirus	Viral infectivity	(Sourisseau et al. 2019)
TEM-1 beta lactamase	<i>S. cerevisiae</i> surface display	Solubility (Fluorescence)	(Klesmith et al. 2017)
Bgl-3	<i>E.coli</i> – Droplet microfluidics	Enzyme activity	(Romero, Tran, and Abate 2015)
β 2AR receptor	Mammalian cell – β 2AR receptor linked barcoded reporter	Receptor activity	(Jones et al. 2019)
M.HaeIII	<i>E. coli</i>	Growth rate, <i>in-vitro</i> selection	(Rockah-Shmuel, Tóth-

Introduction

			Petróczy, and Tawfik 2015)
--	--	--	-------------------------------

GPCR, G protein coupled receptor; Fab, fragment antigen binding; YAP65, yes-associated protein-65; PKA, protein kinase A; PSD95, postsynaptic density protein-95; Hsp90, heat-shock protein 90; Mat α 2, mating type protein- α 2; Pab1, poly(A)binding protein-1; CDR, complementarity-determining region; Bgl-3, Beta-glucosidase; β 2AR, β 2-adrenergic receptor.

Deep mutational scanning as a method to study diverse selection pressures in molecular evolution

Quoting Angela Creager, “When Alan Garen asked Alfred Hershey for his idea of scientific happiness, Hershey replied, “*To have one experiment that works, and keep doing it all the time*”. The first generation of molecular biologists referred to this as *Hershey heaven*” (Creager 2001). Deep mutational scanning can be performed in cells, and can be repeated in slightly different backgrounds to estimate the effect of different conditions on the selection or purge of variants (Fowler and Fields 2014). This ‘deep mutational scanning - Hershey heaven’ could be the right tool to shed additional light on the diversity of selection pressures involved in evolution of proteins. For instance, for a single library, scans can be done at different temperatures to test the prevalence of thermosensitive mutations, in the presence and absence of antibiotics to study the mechanism of resistance evolution, in presence and absence of functional

constraints to map the residues involved in maintaining catalytic fold and function, etc... Some of these maps when build together may help delineate the presence of cryptic selection pressures at work. But for a deep mutational scan to be effective at identifying selection pressures, the system has to be designed to mimic natural conditions as well as possible, emphasising the popular directed evolution maxim 'you get what you select for' (Schmidt-Dannert and Arnold 1999).

Sources of variation in fitness mapping using deep mutational scanning

One of the highly studied constraints on the evolution of proteins is the effect of gene expression levels. Expression level constrains evolution mainly due to two reasons, the fitness cost of misfolded proteins (Ryan et al. 2008), and the fitness cost of mis-interacting proteins (Yang et al. 2012). This fitness cost is linearly correlated to the expression level, since the increased cellular abundance of the misfolded or mis-interacting proteins increases the fitness effect (Geiler-samerotte et al. 2010). These predictions have been systematically tested and are now supported by experimental data (reviewed in (J. Zhang and Yang 2015)). Also, it is an implied assumption that all the variants in a library are expressed at similar levels. But, the variation in expression level due to plasmid copy number variation (Aw and Polizzi 2013) can create artefactual selection or purge of mutations (Jiang et al. 2013; Rockah-Shmuel, Tóth-Petróczy, and Tawfik 2015). Ryan et al. have shown in

Introduction

Saccharomyces cerevisiae that expression of reporter proteins from the genome provides reduced noise in expression levels (Figure 4) (Ryan et al. 2014).

Most studies in fitness mapping utilize mainly *Escherichia coli* or *Saccharomyces cerevisiae* due to their benefits such as high transformation efficiencies, rapid growth rates, well-established manipulation tools, and ability to maintain stable plasmids (Lecomte et al. 2016). However, expression of a foreign gene in a non-native organism and/or from an artificial expression vector can generate differences in selection pressures that are hardly predictable. These can be caused due to a variety of factors such as codon usage, missing chaperones, differences in expression level, physiological conditions, post translational modifications like glycosylation or disulfide bonds (Mate et al. 2011).

In the next two sections, we will present an overview of the model system we will use for performing the deep mutational scan.

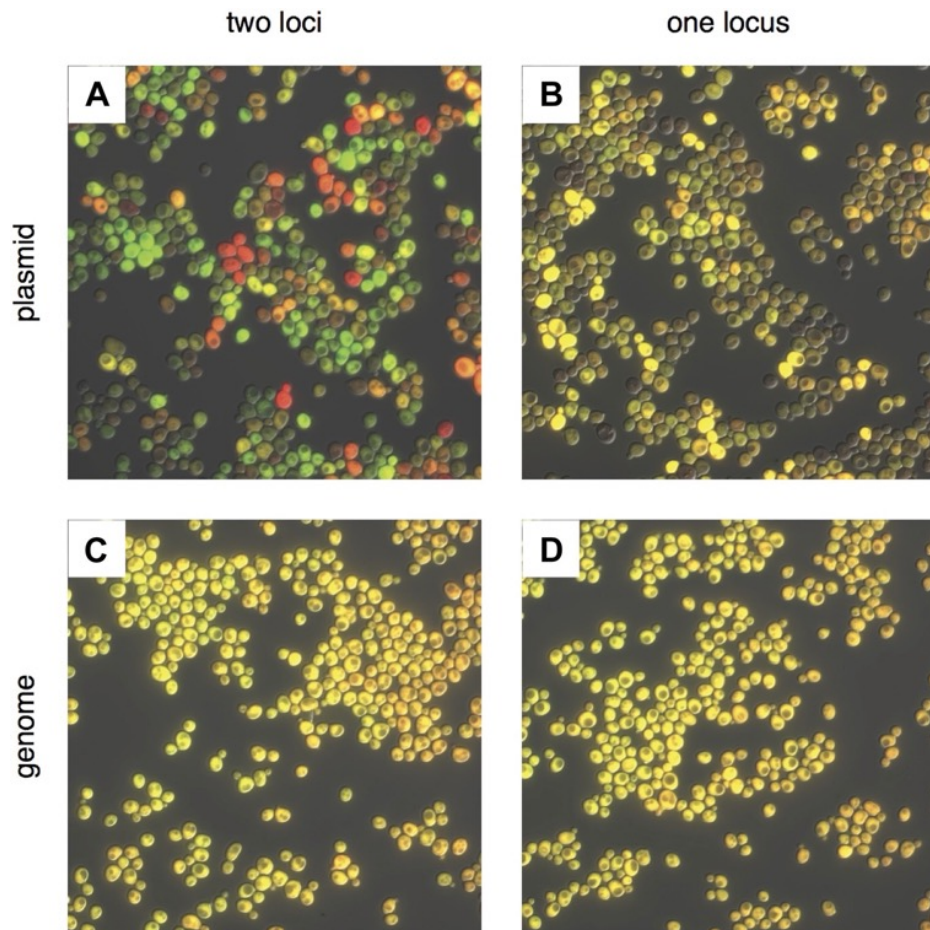


Figure 4. Noise in expression levels of fluorescent proteins cloned in plasmids compared to the genome. Proteins mRuby2 (coloured red) and Venus (coloured green) were expressed with a strong pTDH3 promoter from single cells. (A) proteins expressed from two different CEN plasmids (B) cloned and expressed in tandem on a single CEN plasmid (C) cloned and expressed from two different genomic loci (*leu2* and *ura3*) (D) cloned

Introduction

and expressed from a single genomic locus (*ura3*). Taken from (Ryan et al. 2014).

1.3. *Streptococcus thermophilus* as model organism

S. thermophilus is Gram positive, anaerobic, lactic acid bacterium widely used in the dairy industry for cheese and yogurt production. The U.S. Food and Drug Administration has provided *S. thermophilus* with a “Generally Recognised As Safe” (GRAS) status and the European Food Safety Authority (EFSA) has granted it the “Qualified Presumption of Safety” (QPS) status.

The genome and metabolism of *S. thermophilus* has been studied in detail by P. Hols and co-workers (Bolotin et al. 2004; Hols, Hancy, Fontaine, Grossiord, Prozzi, Leblond-Bourget, Decaris, Bolotin, Delorme, Ehrlich, et al. 2005). As explained by Bolotin et al., *S. thermophilus* contains a 1.8 Mb genome containing 1900 coding sequences. *S. thermophilus* shares 80% of genes (around 1500) with other pathogenic streptococci, but *S. thermophilus* having specialized for the ecological niche of growth in milk, has resulted in loss of all virulence related genes. Interestingly, *S. thermophilus* has a genome which is in evolutionary regression. 10% of its genes are non-functional due to frameshift, nonsense mutations and indels. These are predominantly carbohydrate uptake and metabolism genes which have become unessential for growth in the niche of milk (Bolotin et al. 2004). On the contrary, milk is poor in

free amino acids and *S. thermophilus* encodes well-developed nitrogen metabolism genes to utilize casein and synthesise multiple amino acids. It encodes genes to synthesise multiple amino acids, and can grow well in growth medium supplemented with only glutamine, methionine, leucine, isoleucine, valine and histidine (Letort and Juillard 2001). Additionally, *S. thermophilus* LMD-9 strain also contains a cell envelope anchored PrtS serine protease allowing it to hydrolyse casein from milk (Fernandez-Espla et al. 2000), a set of transporters to import amino acids and peptides and finally a set of intracellular peptidases to further hydrolyse and metabolise casein derived peptides (Hols, Hancy, Fontaine, Grossiord, Prozzi, Leblond-Bourget, Decaris, Bolotin, Delorme, Ehrlich, et al. 2005). *S. thermophilus* requires anaerobic conditions to grow, but it can survive in low concentrations of reactive oxygen species (ROS) due to the presence of an ROS induced stress response pathway (Thibessard et al. 2001).

S. thermophilus LMD-9 is well known for its ability to be naturally competent. As described by Fontaine et al., natural competence is a mechanism of lateral gene transfer by which bacteria can take up naked DNA from the environment and stably integrate it in the genome. Competence in *S. thermophilus* is controlled by the ComRS system (Fontaine, Boutry, De Frahan, Delplace, Fremaux, et al. 2010; Gardan et al. 2009). The mechanism of competence activation by the ComRS is schematized in **Figure 5**.

[illegible]

19

outside the cell, during or after its secretion, it is matured by proteases to release the C-terminal domain, this domain has been named XIP (for σ^x inducing peptide). At a critical extracellular concentration of XIP, it is imported by the Ami/Opp transporter (A-B-C-D-F). Subsequently, XIP interacts with the transcriptional regulator ComR and promotes its dimerization. This ComR-XIP complex binds to the ComR-box located in the promoter of *comX* and *comS* inducing transcription, causing a positive feedback loop on production of ComS. The gene *comX* encodes the alternative sigma factor σ^x , which is responsible for the regulation of late genes important for DNA uptake and recombination. Competence shut-off (red dotted lines) mechanisms are not yet well understood. Taken from (Fontaine et al. 2015)

It has been demonstrated that a synthetic version of the XIP can be added to the growth medium to induce competence (Fontaine et al. 2013a; Gardan et al. 2013a). This allows the highly efficient transformation of *S. thermophilus* with plasmids and linear DNA fragments. Linear DNA fragments can be assembled with ~800 bp homologous arms corresponding to the locus where the gene is to be inserted (Fontaine, Dandoy, Boutry, Delplace, De Frahan, et al. 2010). Dandoy et al. have showed that large fragments upto 15 kb can be easily replaced/inserted to the chromosome with natural transformation (Dandoy et al. 2011).

Introduction

Since working with *S. thermophilus* offers a multitude of benefits including low biosafety requirements, high transformation efficiency and most importantly allowing easy manipulation of the genome. We have thus chosen *S. thermophilus* as the model organism for this study.

S. thermophilus, like all Gram positive bacteria, has a thick peptidoglycan layer in the cell envelope. The pentapeptide L-alanine₁ – D-glutamate₂ – L-lysine₃ – D-alanine₄ – D-alanine₅ is involved in glycan strand crosslinking. Additionally the peptide bridge linking the pentapeptides between L-lysine₃ and D-alanine₄ is composed of two or three L-alanine residues (Delcour et al. 1999). D-alanine is essential for the synthesis of the peptidoglycan layer, it is obtained by racemization of L-alanine by alanine racemase.

1.4. Alanine racemase as a model enzyme

Our primary objective was to study the evolution of a gene from its native genomic locus. For this we will use the **alanine racemase** (AlrSt) of *S. thermophilus* LMD-9. This chapter is a review of information available in the literature about the structure function relationship of alanine racemases. Additionally, we will present a list of important residues in AlrSt as a reference for the rest of the work.

1.4.1. Alanine racemase from *S. thermophilus*

Alanine racemase (Alr, E.C. 5.1.1.1) catalyses the racemization reaction between L-alanine and D-alanine (*Figure 6*) using a covalently-bound pyridoxal-5-phosphate (PLP) cofactor. Alr are ubiquitous in bacteria and absent in most eukaryotes. Appendix Figure 1 shows the distribution of Alr across all domains of life. D-alanine is an essential building block of the peptidoglycan in cell walls and thus it is essential to all bacteria (Lambert and Neuhaus 1972; Walsh 1989). Presence of alanine racemase is ubiquitous in all bacteria and certain fungi. Fungi like *Tolypocladium inflatum* utilize alanine racemase to produce D-alanine which is necessary in the synthesis of cyclosporin A (Hoffmann et al. 1994). Certain bacteria like *Escherichia coli* (Lobocka et al. 1994) or *Pseudomonas aeruginosa* (Strych et al. 2000) encode two alanine racemases, *alr* and *dadX*. In these organisms, deletion of *alr* can be compensated by the second gene which fulfils the metabolic requirement for D-alanine. Moreover, it has been shown that cystathionine beta-lyase (encoded by *metC*) from *E. coli* has a promiscuous alanine racemase activity (Soo et al. 2016), such that upregulation of *metC* can rescue an *alr/dadX* deficient mutant (Kang et al. 2011).

Introduction

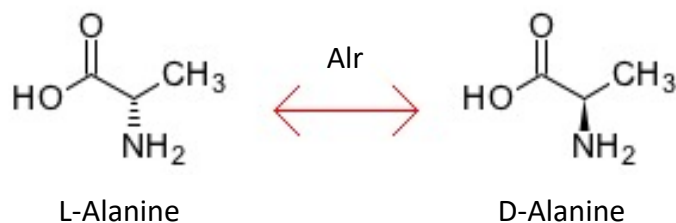


Figure 6. Racemization of L-alanine and D-alanine, catalysed by Alanine racemase (Alr). Modified from KEGG reaction (Kanehisa and Goto 2000)

For our study, we have chosen the native alanine racemase from *S. thermophilus* LMD9 encoded by the gene *alr* and denoted as AlrSt for the rest of this work. AlrSt has 60.93% identity to Alr^{Sp} from *Streptococcus pneumoniae*, 33.78% identity to residues of Alr^{Mt} from *Mycobacterium tuberculosis* (Figure 7) and 47.40 % identity to Alr^{Gs} from *Geobacillus stearothermophilus*. Alr^{Sp} (PDB – 3S46), Alr^{Mt} (PDB – 1XFC) and Alr^{Gs} (PDB – 1BD0) crystal structures are well studied (Im et al. 2011; LeMagueres et al. 2005a; Shaw, Petsko, and Ringe 1997). *S. pneumoniae* and *M. tuberculosis*, as the causative agent of pneumonia and tuberculosis, respectively, and are of high clinical importance. Due to the high homology between these enzymes, results from the study of AlrSt could additionally be relevant for development of Alr inhibitors against these bacteria.

```

ALR_ST 1 MIASLHRP-----TLAKVDLSAISENIEQVVS HIPKQVQTF FAVV KANAYG
ALR_MT 1 MAMTPISQTPGLLAEAMVDLGAEHNVRVIREHAGH-AQLMAVV KADGYG
ALR_SP 1 MKASPHRP-----TKALIH LGAIRQNIQQMGAHIPQGT LK IAVV KANAYG
                                         *

ALR_ST 46 HGAVEVAKH-VSKQVDGFCVSNIDEALELRQAGIEQPILILGVLPDGV-
ALR_MT 50 HGATRVAQTALGAGAAELGVATVDEALALRADGITAPVLAW--LHPECTD
ALR_SP 46 HGAVAVAKA-IQDDVDGFCVSNIDEALELRQAGLSKPILILGVSEIEAV-

ALR_ST 94 -PLAIQENISLTVASLEWLALAQKQGLDLIG--LTCHKVDSGMGRIGVR
ALR_MT 98 FGPALLADVQAVSSLRQLDELLH-AVRRTGRTATVTVKVDITGLNRNGVG
ALR_SP 94 -ALAKYDFILTVAGLEWIQALLDKEVDLTG--LTVHLKIDSGMGRIGFR

h
ALR_ST 141 NLKADNLIAGLK---ALGADV EGIFTHFAT ADE ADDSKKPKRQLSFTDL
ALR_MT 147 PAQFPAMLTALRQAMAEDAVRI RGLMSHMVY A D K P DDS INDVQAQRFTAF
ALR_SP 141 EASEVEQAQDLLQ---QHGVCEGIFTHFAT ADE ESDDYFNAQLERKTI

ALR_ST 188 VDNLTD---RPRLVHASNSATSIWHAATVENTVRIGVVIYGLNPSGSVLE
ALR_MT 197 LAQAREQGVRFVAHLSNSATMARPDLTEDLVRPGIANYGLSPVPAIGD
ALR_SP 188 LASMKE---VPELVHASNSATTLWHVETIFNAVRMGDAMYGLNPSGAVLD

ALR_ST 235 LPYNTQPALSLETALIHVKTI PAGQDVG YGATYTTTAE EVIGTLPIGYAD
ALR_MT 247 M---GLVPAMTVKCAVALVKSIRAGEGVSYGHTWIAPRDTNTALLPIGYAD
ALR_SP 235 LPYDLIPALTLESALVHVKTVPAGACMG YGATYQADS E QVIATVPIGYAD
                                         *

ALR_ST 285 GWTRDLQG-FHVIVDGQLCPIVGVSM DQITVRL---PKVYPLGTPVTLM
ALR_MT 295 GVFRSLGGRLEVLINGRRCPGVGRICMDQFMVDLGPGPLDVAEGDEATLF
ALR_SP 285 GWTRDMQN-FSVLVDGQACPIVGVSM DQITIRI---PKLYPLGTRVTLT

ALR_ST 331 GENGGASITATEVAKRG TINYEVLCLLSDRVPRSYD-----
ALR_MT 345 GPGIRGEPTAQDWADLVGTIHYEVVTS PRGRITRTYREAENR
ALR_SP 331 GSNQDKEITATQVATYRV TINYEVVCLLSDRIPREY-----

```

Figure 7. Multiple sequence alignment showing Alr(s) from *S. thermophilus* (ST) (STER_RS08315) (GenBank accession number: Q03IX8), *M. tuberculosis* (MT) strain ATCC 25618 / H37Rv (GenBank accession number: P9WQA9) and *S. pneumoniae* (SP) serotype 4 strain ATCC BAA-334 / TIGR4 (GenBank accession number: P0A2W8). The active site lysine and tyrosine are shown with an asterisk(*), the core residues of the active

Introduction

site are shown in red, the residues between the active site and the entryway are shown in blue and the outer residues of the active site gateway are shown in green. T-Coffee (Di Tommaso et al. 2011) was used to build the multiple sequence alignment.

1.4.1.1. Structure of alanine racemase

The crystal structure of alanine racemases from many Gram positive and negative bacteria have been solved and deposited to PDB. Structure of alanine racemases from *Geobacillus stearothermophilus* (Alr^{Gs}), *P. aeruginosa*, *Streptomyces lavendulae*, *Mycobacterium tuberculosis* (Alr^{Mt}), *Bacillus anthracis*, *E. coli*, *Enterococcus faecalis*, *Bartonella henselae*, *Oenococcus oeni*, *Pseudomonas fluorescens*, *Actinobacillus succinogenes*, *Corynebacterium glutamicum*, *Streptococcus pneumoniae* (Alr^{Sp}) and *Staphylococcus aureus* are available on PDB. In all these cases, Alr share a common homo-dimeric fold formed by head-to-tail association of two monomers (Figure 7B). The active site is located in a cleft at the interface between the two subunits. Monomeric Alr have also been described in *Thermus thermophilus* (Inagaki et al. 2000), *Salmonella typhimurium* (Esaki and Walsh 1986) and *Shigella* species (Yokoigawa et al. 2001) but their similarity to dimeric Alrs is very low. A subunit of Alr consists of an eight-stranded α/β barrel domain at the N-terminal and an extended β -strand domain at the C-terminal (Figure 7A). The active site is formed between the α/β barrel domain from one

monomer and the extended β -strand domain from the second monomer. The cofactor PLP is covalently attached to the Lys40 residue situated inside the α/β barrel domain (Figure 8). The lysine bound PLP molecule forms a N'-pyridoxyl-lysine-5'-monophosphate or LLP residue where the aldehyde function of PLP is converted into an aldimine (Figure 8B).

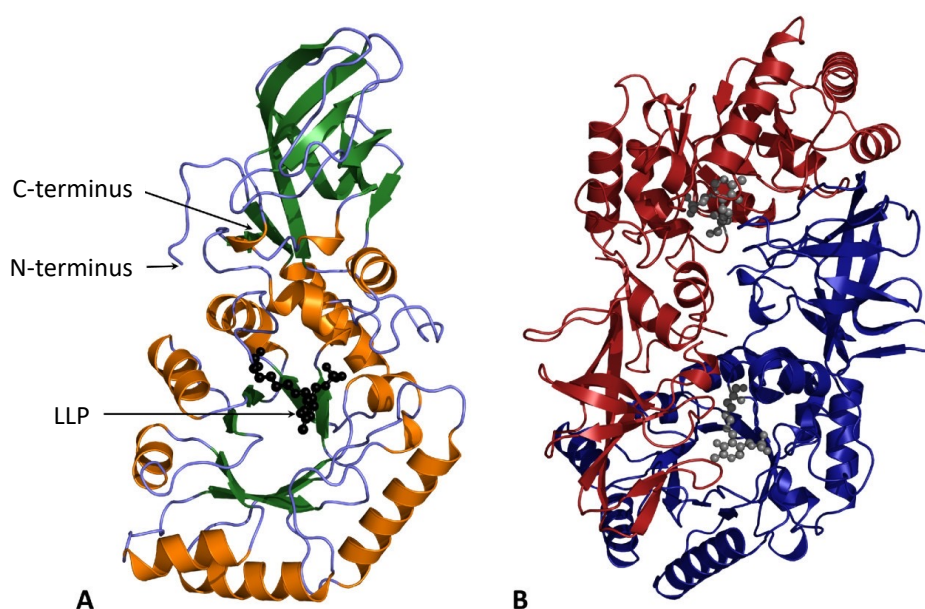


Figure 8. (A) Ribbon diagram of one monomer of Alr^{Sp}. The alpha helices are coloured orange and the beta sheets are coloured green. **(B)** Ribbon diagram of both Alr^{Sp} dimer. One of the monomers is coloured red and the other is coloured blue. The LLP residue, depicted as black or grey balls, is located at the interior of the α/β barrel domain of the active site. (PDB – 3S46) Taken from (Im et al. 2011).

Introduction

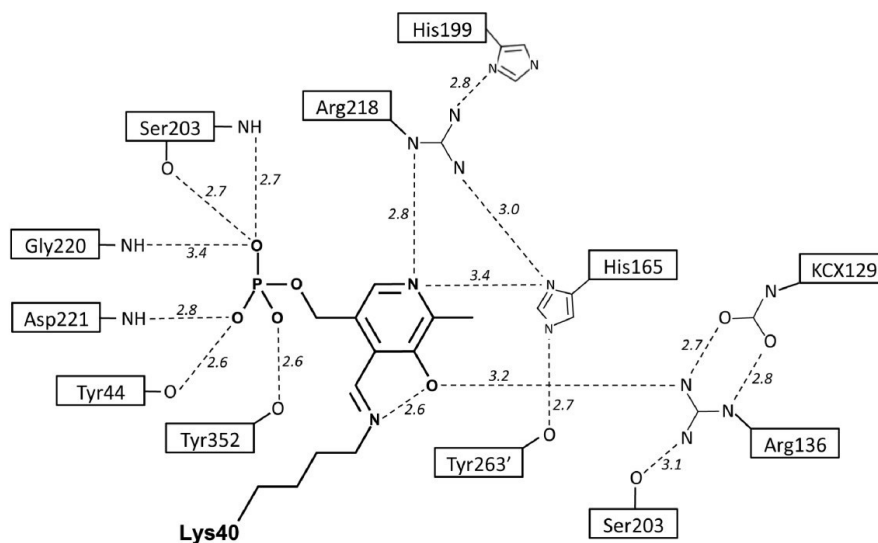


Figure 9. Active site of Alr^{SP}. Schematic representation of the interactions around PLP in the active site. Dotted lines represent the distances between the atoms. The active site lysine is displayed in bold as an internal aldimine. Taken from (Im et al. 2011)

Active site and conserved residues

There are two catalytic residues in the active site of Alr, Lys40 and Tyr263' (primed numbers indicate residues from the second monomer). These two residues are strictly conserved across most Alr homologs. The PLP binding site motif near the N-terminus (AVVK₄₀ANAYGHG) is also conserved.

The PLP molecule is stabilized by group of essential residues in its neighbourhood. As already mentioned, PLP is bound by a covalent double

bond forming the internal aldimine with Lys40 (Noda et al. 2004; Watanabe et al. 2002). PLP is further stabilized by hydrogen bonds with side chains of residues Tyr44, Arg136, His165, Ser203, Arg218, Tyr352, with the main chains of residues Ser203, Gly220, Asp221, and additionally with His199 and Tyr263' (Figure 9). The carboxylate group of the substrate alanine is stabilized in the structure by hydrogen bonding with Arg136; Ser309' and a carbamylated Lys129 helps position the Arg136 by hydrogen bonding (Im et al. 2011; LeMagueres et al. 2003, 2005b; Morollo, Petsko, and Ringe 1999; Shaw, Petsko, and Ringe 1997).

Active site entryway

The LLP group is inside the α/β barrel domain and the substrate alanine molecule has to pass through a constricted entryway to reach the aldimine reactive center. This entryway is conserved in most alanine racemases. It is composed of three layers of residues – inner, middle and outer. The inner and middle layers are strictly conserved (Im et al. 2011; LeMagueres et al. 2005b) while the outer layer residues can vary (Table 3).

Table 3. Residues conserved in forming the active site entryway in Alr^{Mt}, DadX^{Pa} (DadX from *P. aeruginosa*), Alr^{Gs}, Alr^{Sp} and AlrSt

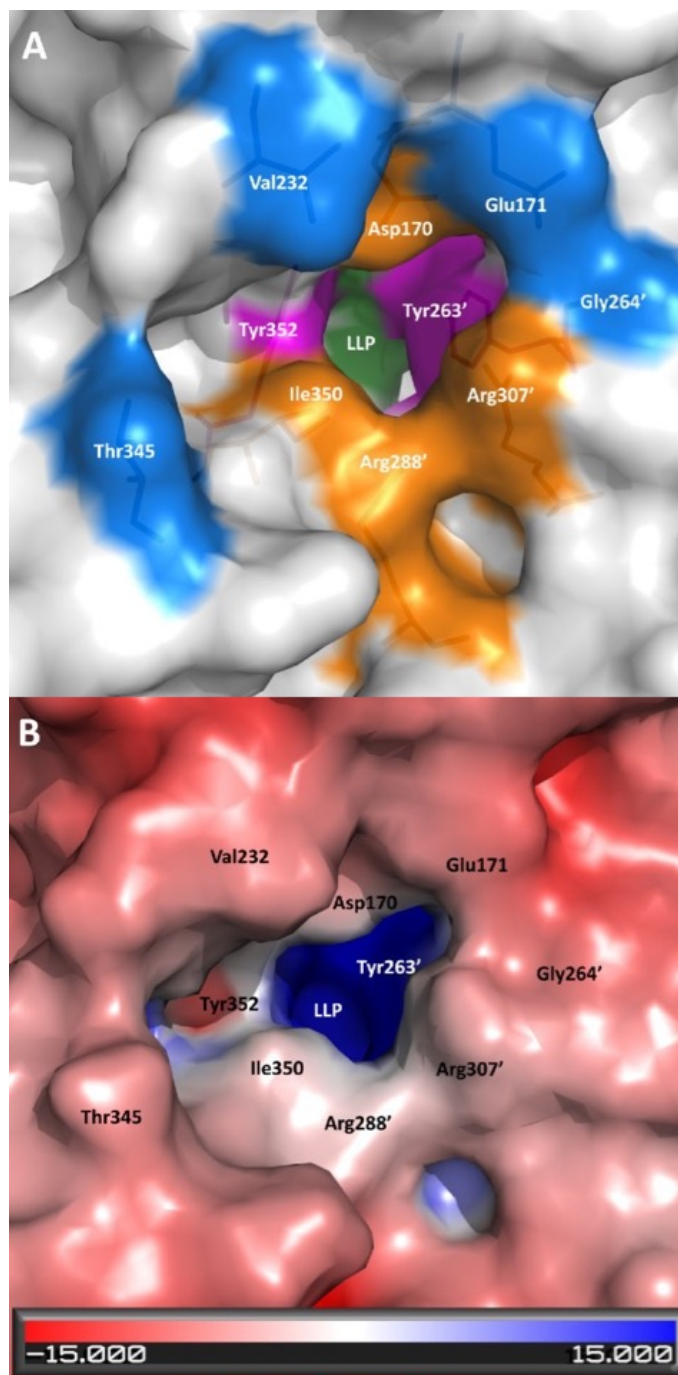
Position	Alr ^{Mt}	DadX ^{Pa}	Alr ^{Gs}	Alr ^{Sp}	<u>AlrSt</u>
Inner layer					

Introduction

1	Tyr271'	Tyr253'	Tyr265'	Tyr263'	Tyr263'
2	Tyr364	Tyr341	Tyr354	Tyr352	Tyr352
3	Tyr290'	Tyr272'	Tyr284'	Tyr282'	Tyr282'
4	Ala176	Ala162	Ala170	Ala169	Ala169
Middle layer					
1	Arg316'	Arg298'	Arg309'	Arg316'	Arg307'
2	Ile362	Ile339	Ile352	Ile350	Ile350
3	Arg296'	Arg278'	Arg290'	Arg288'	Arg288'
4	Asp177	Asp163	Asp171	Asp170	Asp170
Outer layer					
1	Asp359	Ala334	Arg347	Thr345	Glu345
2	Lys180	Glu164	Glu172	Glu171	Glu171
3	Leu244	Ala222	Ile233	Val232	Val232
4	Gly274'	Gly254'	Gly266'	Gly264'	Gly264'

AlrSt is underlined for emphasis. Modified from (LeMagueres et al. 2005b).

Negatively charged Asp170 and Glu171, along with positively charged Arg288' and Arg307 from the middle layer have been proposed to have an important role in controlling the proper substrate orientation for entry into the active site tunnel (D. Wu et al. 2008).



Introduction

Figure 10. The entryway into Alr^{Sp} active site represented as a molecular surface. (A) The three layers of the active site entryway, inner layer residues coloured in pink, middle layer residues coloured in orange and outer layer residues coloured in blue. AlrSt encodes the same inner and middle layer residues, only exception is the outer layer Glu345 (AlrSt) instead of Thr345 (Alr^{Sp}). **(B)** Representation of the electrostatic potential of the surface of the entryway (PDB – 3S46). Taken from (Im et al. 2011).

Table 4. List of all the important active site residues in AlrSt

Active site residues	
Alr St position	Annotated function
K40	Catalytic residue
Tyr263'	Catalytic residue/Inner entryway/Stabilize PLP cofactor
Tyr352	Substrate specificity/Inner entryway
Tyr282'	Inner entryway
Ala169	Inner entryway
Arg307'	Middle entryway
Ile350	Middle entryway
Arg288'	Middle entryway
Asp170	Middle entryway
Tyr44	Stabilize PLP cofactor
Arg136	Stabilize PLP cofactor
His165	Stabilize PLP cofactor
Ser203	Stabilize PLP cofactor
Arg218	Stabilize PLP cofactor

Tyr352	Stabilize PLP cofactor
Gly220	Stabilize PLP cofactor
Asp221	Stabilize PLP cofactor
His199	Stabilize PLP cofactor
Met310'	Stabilize reaction intermediates
Lys129	Stabilize Arg136
Ser309'	Stabilize Arg136

Primed numbers denote residues from the second monomer (Im et al. 2011; LeMagueres et al. 2003, 2005b; Morollo, Petsko, and Ringe 1999; Shaw, Petsko, and Ringe 1997)

Dimer interface

Studies with Alr^{Pa} and DadX^{Ec} show that dimerization is essential for catalytic activity (Strych and Benedik 2002). The active site, the entryway and the PLP-binding pocket is composed by residues from both monomers. We used a swiss model predicted structure of AlrSt as input for PISA for identifying intermonomer residues (Krissinel and Henrick 2007). Across the interface, there are 37 hydrogen bonds and 16 salt bridges, while there are no disulphide bridges or covalent bonds. 19% of the total surface area of both monomers are involved in the intermonomer interactions, similar to observations in Alr^{Gs} and Alr^{Ef} (Im et al. 2011). Im et. al. describe a cluster of hydrogen-bonded water molecules forming an ordered pentagonal ring in Alr^{Sp} making hydrogen bonds with residues involved in the interface (Asp170, Tyr263', Tyr282', Tyr288', Arg307', Tyr352). These water molecules have been proposed (1) to act as a mobile

Introduction

network of hydrogen bonds stabilizing the dimer interface, and (2) to supply the active site with water molecules allowing for rapid equilibration of active site protons with the solvent (Mustata and Briggs 2004; Spies and Toney 2003).

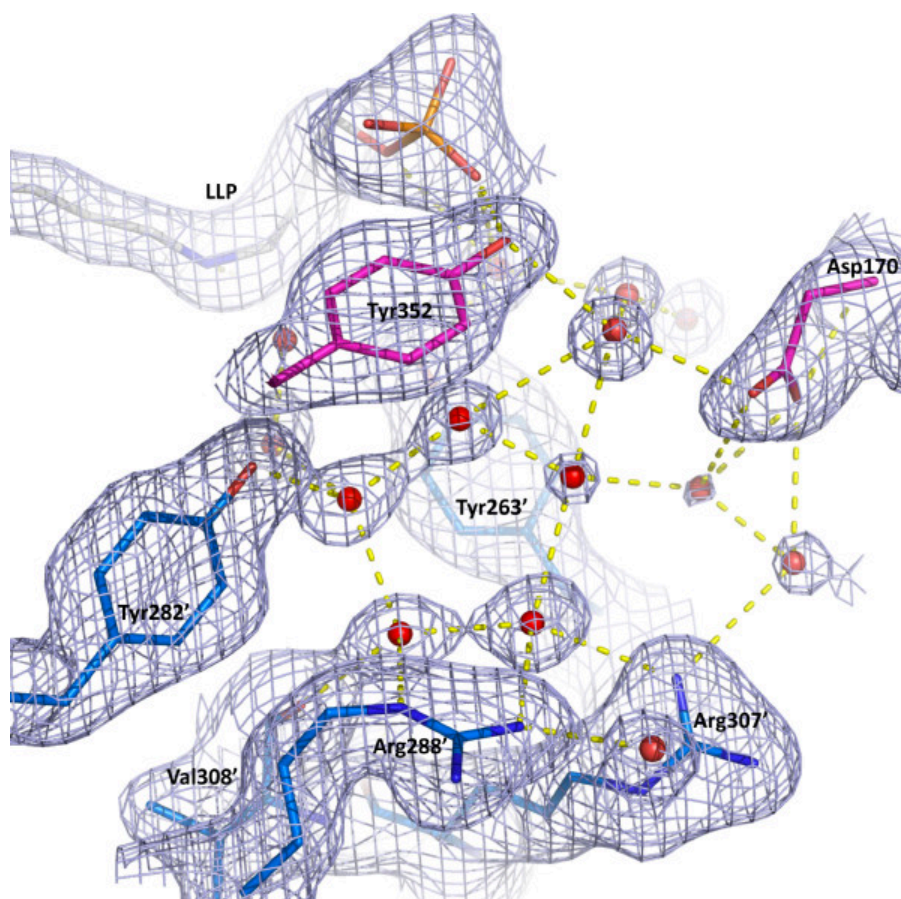


Figure 11. Pentagonal ring waters near the active site in Alr^{Sp}. The mesh represents the electron density 2Fo-Fc map. Residues are represented as sticks, hydrogen bonds as dashed yellow lines and water molecules are represented as red spheres. Residues from the first and second monomer

are shown in blue and pink respectively and residue numbers are denoted by prime (') symbols. The LLP moiety is represented in grey. (PDB – 3S46). Taken from (Im et al. 2011).

1.4.1.2. Reaction mechanism

The reaction mechanism of Alr has been studied in detail over the decade. Alr shows a different reaction mechanism than most other PLP-containing enzymes. Usually the pyridine nitrogen of PLP is protonated, making hydrogen-bonds either with a negatively charged residue (Asp or Glu) in aspartate aminotransferase or a polar residue (Ser or Thr) in tryptophan synthase family. In this case, the pyridinium ion acts as an effective electron sink which stabilizes the carbanion intermediate by forming a quinonoid species. However, Alr are unique among the PLP-containing enzymes as the pyridine nitrogen of PLP is unprotonated. This is because of the presence of a positively charged Arg residue which prevents the formation of the cationic pyridinium ion. In AlrSt, the equivalent position is occupied by Arg218. The Alr reaction probably follows without the formation of the quinonoid intermediate.

Here, we will provide a general summary of the reaction mechanism taken from (Holliday et al. n.d.). The scheme is depicted in Figure 12 and summarizes the conclusions of several independent studies on different Alr homologs (LeMagueres et al. 2005b; Dan T. Major, Nam, and Gao 2006; Dan Thomas Major and Gao 2006; Ondrechen, Briggs, and McCammon 2001; Shaw, Petsko, and Ringe 1997; Spies and Toney 2007;

Introduction

Stamper, Morollo, and Ringe 1998; Sun and Toney 1999; Watababe et al. 1999; Watanabe, Kurokawa, et al. 1999; Watanabe, Yoshimura, et al. 1999; Watanabe et al. 2002).

Initially the PLP cofactor is bound to Lys40 with an aldimine linkage.

1. In the first step, the amine group of L-alanine substrate attacks the aldimine carbon in a nucleophilic addition, Lys40 deprotonates the formed ammonium, and the resulting secondary amine eliminates the bound lysine, while forming a new aldimine between alanine and PLP.

Next, anionic Tyr263' deprotonates the alpha carbon of PLP-alanine intermediate, resulting in a carbanionic intermediate. This is the rate determining step. This carbanion intermediate is probably stabilized by strong solvation effect and is postulated because the quinonoid intermediate cannot be formed.

2. The ammonium of Lys40 reprotonates the carbanion from the other side.
3. In a reverse step compared to step 1, the amine of Lys40 initiates an elimination of D-alanine product and reforms the internal aldimine with PLP.

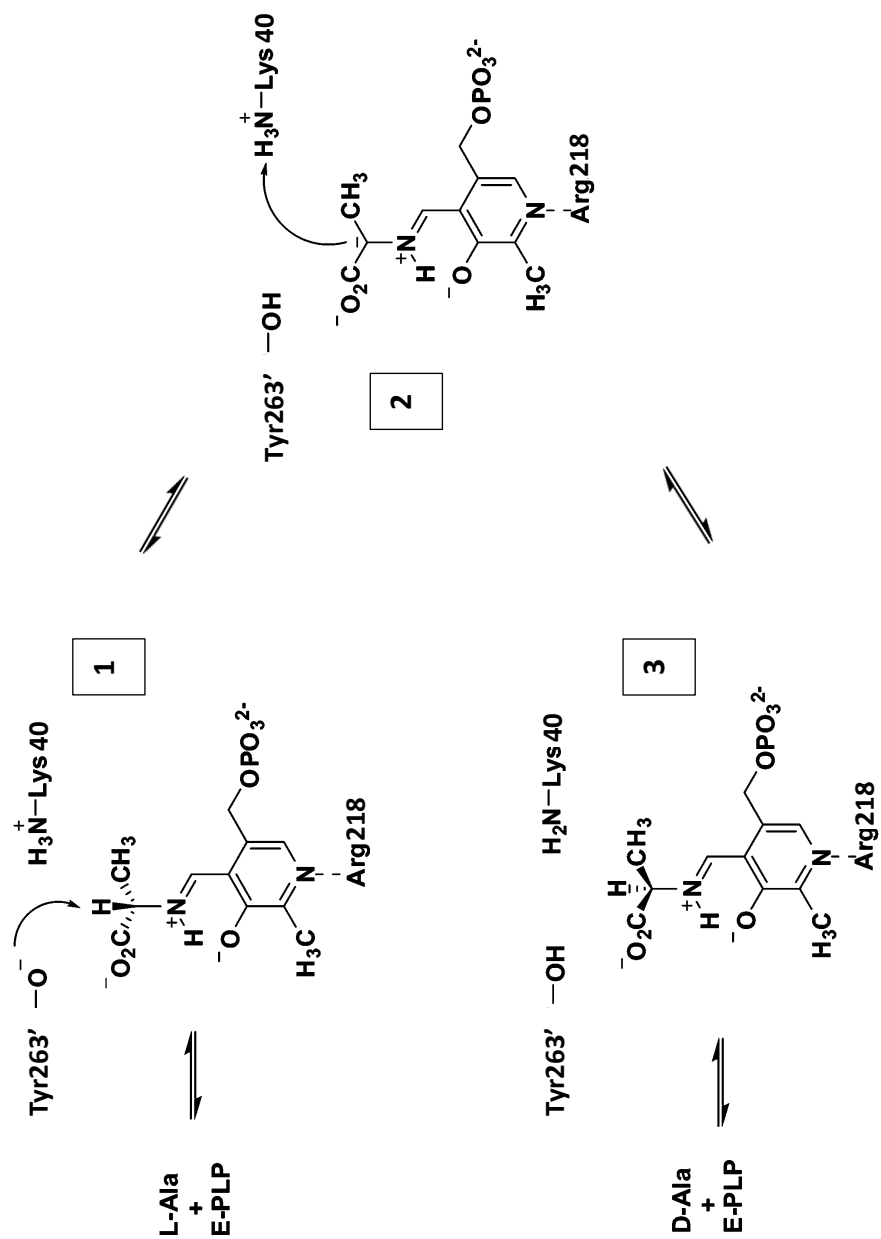


Figure 12. General reaction mechanism of alanine racemase. Modified from (Dan Thomas Major and Gao 2006).

2.

Aim of the project and strategy

The primary aim of this project is to study the selection pressures governing molecular evolution of proteins. Broadly, we are asking the question “*what are the different selection pressures acting on the evolution of a protein?*”. For this, we present an experimental method to study the evolution of a library of Alr variants under two different conditions

(1) Growth in media without D-alanine supplementation

To study how a constraint on maintaining native Alr activity limit the genotypic space available to mutants.

(2) Growth in media with D-alanine supplementation

In media with supplemented D-alanine, maintaining the native function of Alr is no longer under selection. In natural environments, enzyme functions are not always required, so we wish to study how

an enzyme evolves in the absence of selection pressure for function. We are asking the question if there are still predominant selection pressures that act to select certain mutants? Can we identify some of these selection pressures?

Our experimental setup offers several advantages for this study. First, in our model organism *S. thermophilus*, variants can be cloned to the native chromosomal locus with high transformation efficiency. This offers us two benefits (1) consistent and natural gene expression levels (2) ability to create large variant library sizes. Second, several features of our model enzyme Alr make it interesting and suitable for this study. (1) Alr is a well characterized enzyme. Thus, we already have a wealth of information regarding the reaction mechanism and the effect of mutations on enzyme activity. (2) Deletion of Alr activity creates D-alanine auxotrophy in all previously reported bacteria. Thus bacteria expressing active Alr variants are easily selectable in media devoid of D-alanine. (3) The 3-dimensional structures of homologous Alr from *Streptococcus pneumoniae* (PDB: 3S46), *Enterococcus faecalis* (PDB: 3E5P) and *Geobacillus stearothermophilus* (PDB: 1BD0) are already known. (4) Alr is a dimeric enzyme and we would like to compare selective pressures acting on buried residues within a subunit or at the interface. (5) Alr is known to be a highly promiscuous enzyme (Di Salvo et al. 2013; Soo et al. 2016; Yow et al. 2003) and we would like to explore possible selective pressures that are acting on these promiscuous activities, for example mutations that are

increasing or decreasing a promiscuous activity because it is beneficial or detrimental to fitness. In the absence of pressure for Alr activity, our experimental setup may allow to detect mutations that are modifying promiscuous activities and affecting the fitness of the expression host. Hence, we expect to use this system to generate a precise scan of the effect of mutations in alanine racemase to the growth of *S. thermophilus*, and derive inferences regarding subtle selection pressures acting on the evolution of a protein. The results are described over Chapter 3.1 and Chapter 3.2.

With the lockdown during the COVID outbreak in 2020, we also decided to use *S. thermophilus* to develop a method to screen libraries of millions of peptide molecules cyclised *in-vivo* by split-intein circularized ligation of peptides and proteins (SICLOPPS) and search for inhibitors of the SARS-CoV-2 Papain-like protease (PLpro). For this, our strategy was to first create a PLpro induced conditional D-alanine auxotroph mutant. Subsequently we used this mutant to select for peptidic inhibitors of PLpro.

The benefits of our experimental system are (1) *S. thermophilus* permits ease of chromosomal editing and high efficiency of transformation. (2) As described for the primary aim, *S. thermophilus* Alr mutants have a selectable phenotype, allowing easy selection of auxotrophic and prototrophic mutants. (3) Finally, the peptide library generation method known as SICLOPPS is ideal since it allows for

inexpensive *in-vivo* synthesis and screening of billions of genetically encoded randomized peptide molecules.

We expect to use this system to select peptidic inhibitors of PLpro, potentially leading to therapeutic molecules to treat SARS-CoV-2 infections. The results are described in Chapter 3.3.

3.

Results

3.1. *Streptococcus thermophilus* as a suitable host for constructing large scarless variant libraries cloned and expressed from the chromosome

Jason Baby Chirakadavil, Pascal Hols, Patrice Soumillion

Biochemistry and Genetics of Microorganisms (BGM), Louvain Institute of Biomolecular Science and Technology (LIBST), Université catholique de Louvain, Place Croix du Sud, 4-5 box L7.07.06, 1348 Louvain-La-Neuve, Belgium.

S. thermophilus as a suitable host for constructing large scarless variant libraries cloned and expressed from the chromosome

3.1.1. Abstract

The ability to create enzyme variant libraries with billions of mutants has many applications in directed evolution and synthetic biology. However, mutations can create fitness defects due to toxic or misfolded variants. Moreover, this effect on fitness is intensified in highly expressed genes owing to a higher dose of toxicity, misfolded protein aggregation etc. (Ryan et al. 2008). However, most current directed evolution approaches often neglect the cost of overexpression from synthetic plasmid constructs (Aw and Polizzi 2013). Variants evolving from the chromosome mimics the natural evolutionary process better, moreover it offers stable and similar protein dosage level for all variants in the library. In this work, we have developed a simple two-step method for highly efficient and scarless genetic modification of the chromosome in *S. thermophilus* LMD-9. For this, we have taken advantage of the high natural competence of this strain and use of a *cat-oroP* cassette for simple double step cloning by selection/counter selection. Using this system, we have successfully deleted and inserted genes at multiple loci. Finally, we have demonstrated the use of the high efficiency of transformation (upto 10^9 transformants) for creating large gene variant libraries, by creating an error prone PCR library of 10^8 variants of alanine racemase (*alr*) mutants cloned at its native chromosomal locus. Additionally, since natural transformation in *S. thermophilus* allows transformation of up to 10% of viable cells, we deleted *lysA* gene from the cells encoding the library of *alr*

Results Chapter 3.1

without significant loss of diversity demonstrating that it is possible to simultaneously modify the genetic background of all the variants in a library.

3.1.2. Introduction

Study of molecular evolution using *in-vivo* evolutionary systems allows us to map the phenotypic effects of mutations. Most current studies utilize extra-chromosomal elements like plasmids or bacteriophages for ease of genetic manipulation. However, it has been shown that there can be considerable variation of expression levels from genes cloned at extra-chromosomal vectors. Stochastic high expression can mask deleterious effects of mutations since many enzymes display efficiencies much higher than the threshold for biological activity (Newton, Arcus, and Patrick 2015), or exhibit phenotypes which would be unrealistic in native expression conditions. Additionally, it has also been shown that the fitness effect of mutations are proportional to the dosage of the mutant protein (Geiler-samerotte et al. 2010; Jiang et al. 2013; Mehlhoff et al. 2020). Studies with Hsp90 have shown that mutations which are tolerated at lower expression levels display deleterious phenotypes at higher expression level (James, L. C., Tawfik 2001; Jiang et al. 2013; McLoughlin and Copley 2008; Rockah-Shmuel, Tóth-Petróczy, and Tawfik 2015; Zhou et al. 2012) Similarly, studies have demonstrated that changes in gene expression level can also create biologically relevant

S. thermophilus as a suitable host for constructing large scarless variant libraries cloned and expressed from the chromosome adaptations (Soo, Hanson-Manful, and Patrick 2011). Ideally, to accurately map the phenotypic effects of mutations in the genotype, the variation in phenotype should only originate from the variation in the genotype (Zhou et al. 2012). Previous studies have tried to tackle this problem in multiple ways, such as by drastically reducing expression levels by means of low copy number plasmids (Jiang et al. 2013; Rockah-Shmuel, Tóth-Petróczy, and Tawfik 2015) or by integrating the mutant gene to the chromosomal locus. However, existing methods to clone libraries in the chromosome are cumbersome and suffer from low efficiency of transformation. Ryan et al. have previously demonstrated the benefits of chromosomally-integrated variant libraries, but the size of the library reported by them was relatively small (1.6×10^3) (Ryan et al. 2014). One of the reason for lack of many alternatives is that most work has been limited to *Escherichia coli*, the standard workhorse of molecular biology, which unfortunately lacks a highly efficient DNA uptake and recombination system. (Sinha and Redfield 2012).

S. thermophilus is a Gram-positive bacterium, which is used extensively for cheese and yogurt production. Natural transformation in *S. thermophilus* LMD-9 is very well studied (Fontaine et al. 2013a; Fontaine, Boutry, De Frahan, Delplace, Fremaux, et al. 2010; Gardan et al. 2009, 2013a). It relies on the secretion and reimportation of a peptide pheromone ComS to activate a cascade of regulators resulting in the induction of DNA uptake and recombination genes. A peptide which acts

Results Chapter 3.1

as an inducer of natural competence can be added to the medium externally (Fontaine, Dandoy, Boutry, Delplace, De Frahan, et al. 2010; Fontaine et al. 2013a; Gardan et al. 2013a). Fontaine et al. have previously reported a system for markerless modification of the chromosome in *S. thermophilus*, but it involves multiple cloning and screening steps which limit its use for generation of a library. It has been previously demonstrated that the gene *oroP* from *Lactococcus lactis* could be utilized as a counter-selection marker in *Escherichia coli* and *Bacillus subtilis* (Defoor, Kryger, and Martinussen 2007). This gene encodes an orotate transporter which can transport a pyrimidine analogue 5-fluoorotate (5-FOA) into the cell. Intracellularly, 5-FOA can be metabolized by the enzymes of the pyrimidine metabolism resulting in the toxic metabolite 5-fluorodeoxyuridine monophosphate, which covalently binds to thymidylate synthase and prevents formation of dTMP. Thus *oroP* can be utilized as a counter-selection marker by conferring toxicity for 5-FOA (Defoor, Kryger, and Martinussen 2007). Mignolet et al. demonstrated that *oroP* based counter-selection can be used in *Streptococcus salivarius* (Mignolet et al. 2018).

In this work, we introduce an elegant solution to chromosomal library construction by using the naturally competent bacterium *S. thermophilus*. We developed a *cat-oroP* bicistronic cassette which we used iteratively to create multiple scarless chromosomal modifications with exceptionally high efficiency. As a proof-of-concept, we apply the

S. thermophilus as a suitable host for constructing large scarless variant libraries cloned and expressed from the chromosome strategy to generate a large library of genes directly at its native locus. Additionally, we also present an application of this method in situations where the genetic background of an already existing library needs to be modified.

3.1.3. Results

oroP* can be used as a counter selection marker in *S. thermophilus

To verify if *oroP* can be utilized as a counter selection marker in *S. thermophilus*, we first evaluated the toxicity of 5-FOA on wild-type LMD-9 strain. Similar number of colonies grew in plates containing increasing concentrations of 5-FOA (up to 3.5 mg/ml), although colony sizes were smaller at concentrations higher than 2.5 mg/ml, indicating minor toxicity at high concentrations. Based on these results, we constructed a cassette containing *oroP* and *cat* gene (chloramphenicol acetyl transferase) to replace the *leuB* gene (STER_RS05775) encoding isopropyl-malate dehydrogenase involved in leucine biosynthesis

For each gene replaced, the coding sequence of the gene was entirely replaced with the cassette, in the same orientation as the native gene. The native promoter and Shine-Dalgarno sequence were left untouched. These cassettes were used as donor DNA for natural transformation of the wild-type LMD-9 strain (**Figure 1. 1 A and B**). Chloramphenicol resistant transformants were selected on media supplemented with 5 µg/ml chloramphenicol (Chl) to create the strains

Results Chapter 3.1

$\Delta alr :: P_{32-cat} P_{SP-oroP}$ and $\Delta leuB :: P_{32-cat} P_{SP-oroP}$. A single colony was picked, regrown and 10-fold dilutions were spotted onto solid growth media to verify auxotrophy. The Δalr strain displayed D-alanine auxotrophy and $\Delta leuB$ strain displayed leucine auxotrophy (**Figure 1. 1 C**). For subsequent experiments, when necessary, for growth D-alanine or leucine were added when not already present in the growth medium.

Next, we decided to estimate the growth of these mutant strains in media supplemented with 5-FOA. The strains were grown with addition of different concentrations of 5-FOA. We observed a very slow growth of the mutants in presence of 5-FOA (1.5mg/ml) (**Figure 1. 2 A**) while the wild-type strain showed only a slight reduction in growth (**Figure 1. 2 B**). Similar results were observed for the $\Delta alr :: P_{32-cat} P_{SP-oroP}$ and $\Delta leuB :: P_{32-cat} P_{SP-oroP}$ strains on solid media, where no colonies were observed in media supplemented with 1.5 mg/ml of 5-FOA.

Interestingly, we observe that at lower concentrations of 5-FOA (0.5mg/ml) no colonies are observed in the $\Delta leuB :: P_{32-cat} P_{SP-oroP}$ strain however a few colonies are still observed in the $\Delta alr :: P_{32-cat} P_{SP-oroP}$ strain. Although at higher concentrations of 5-FOA (1.5 mg/ml) we observe no colonies in both strains (**Figure 1. 3 A**). We hypothesise that that the concentration of 5-FOA required to create a growth defect seems to be inversely related to the expression level of the *oroP* gene. In order to verify this, we characterized the expression level of the *alr* and *leuB* chromosomal locus. *alr* is expressed under a constitutive promoter

S. thermophilus as a suitable host for constructing large scarless variant libraries cloned and expressed from the chromosome (discussed in detail in Section 3.3 “*Alr* is expressed constitutively in *S. thermophilus*”) and *leuB* is part of the leucine operon. The native *alr* and *leuB* gene coding sequence were replaced by the *luxAB* luciferase reporter gene (**Figure 1. 3 B**), and luminescence was measured. The *leuB* gene is expressed at a 2.2-fold higher level than the native *alr* gene (**Figure 1. 3 C**). This was also observed previously (Solem et al. 2008) and we believe that optimization of *oroP* for higher expression in *S. thermophilus* could allow use of much lower concentrations of (the quite expensive) 5-FOA for the counter-selection.

Results Chapter 3.1

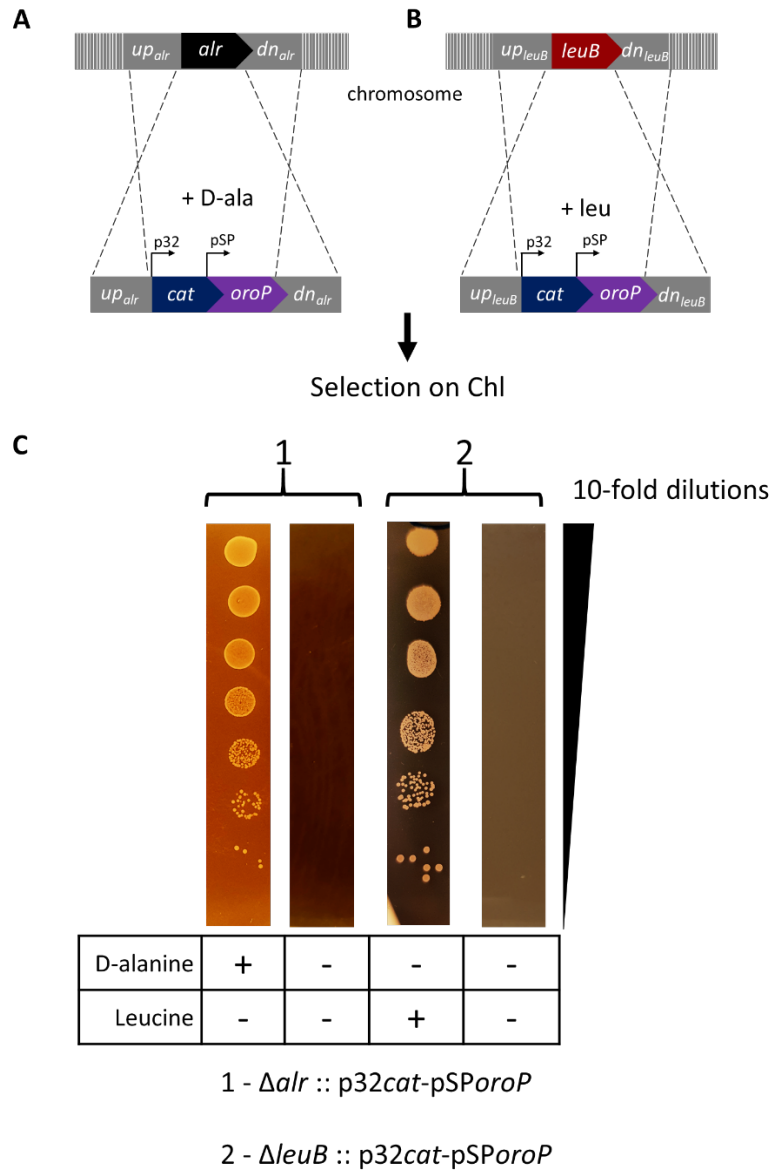


Figure 1. 1. Construction of the Δalr and $\Delta leuB$ strains and auxotrophy phenotype (A) Schematic representation of the cassette for replacing *alr* gene with P_{32} -cat P_{SP} -oroP. (B) Schematic representation of the cassette

S. thermophilus as a suitable host for constructing large scarless variant libraries cloned and expressed from the chromosome for replacing *leuB* gene with P_{32} -cat P_{SP} -oroP. Vertical white lines on the bar are used to represent the chromosomal context of the cloning. The promoters are indicated by black arrows. **(C)** Auxotrophy phenotype of the mutants. Dilutions of Δalr strain were spotted onto M17 growth media plates with and without supplementation of D-alanine, similarly $\Delta leuB$ strain were spotted onto CDM growth media plates with and without supplementation of leucine.

Results Chapter 3.1

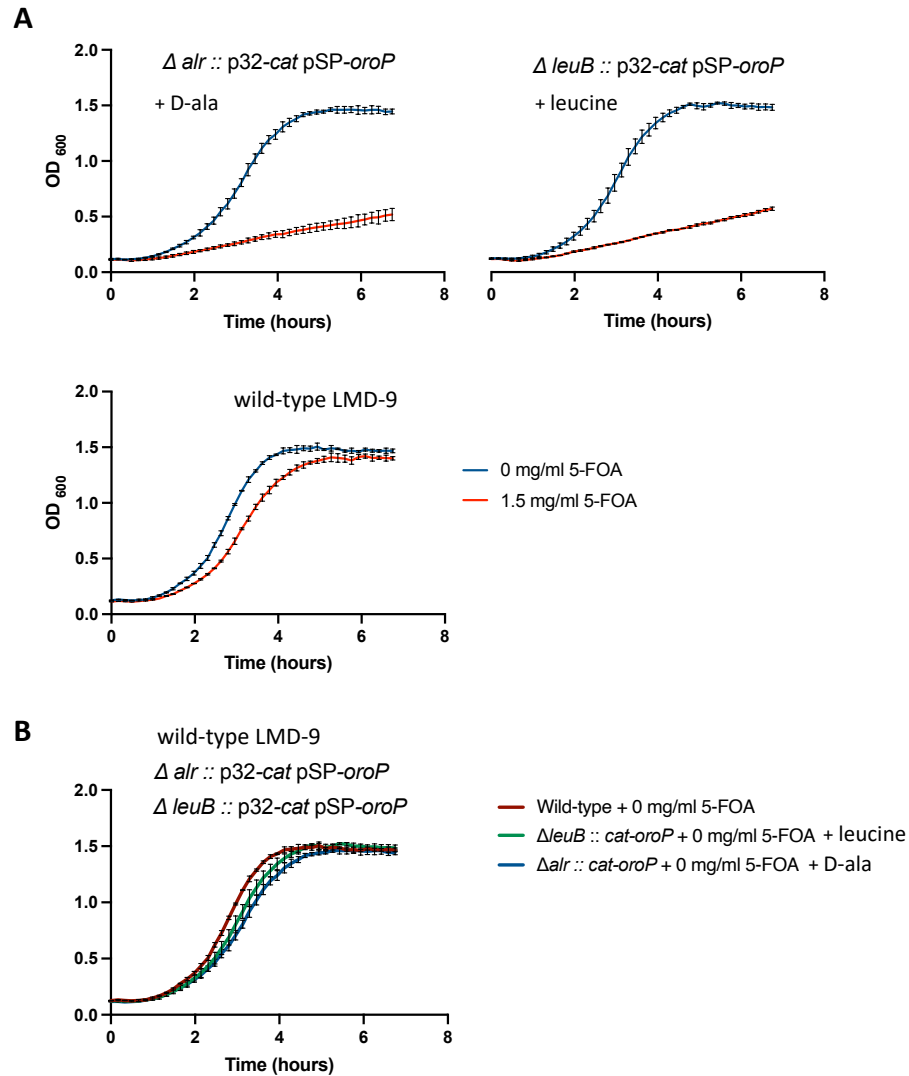


Figure 1. 2. Effect of 5-FOA on growth of *oroP* encoding strains. (A)

Growth expressed as the OD₆₀₀ values measured over time for the wild-type LMD-9 strain versus the Δalr and $\Delta leuB$ mutant strains in media supplemented with 5-FOA. **(B)** OD₆₀₀ values of wild-type, Δalr and $\Delta leuB$

S. thermophilus as a suitable host for constructing large scarless variant libraries cloned and expressed from the chromosome in media without 5-FOA supplementation shows a minor retardation in growth. Growth was measured in triplicates at 10 minutes intervals on a Tecan Infinite 200 PRO series (Tecan, Männedorf, Switzerland). Mean values calculated from three independent replicates are represented here with error bars

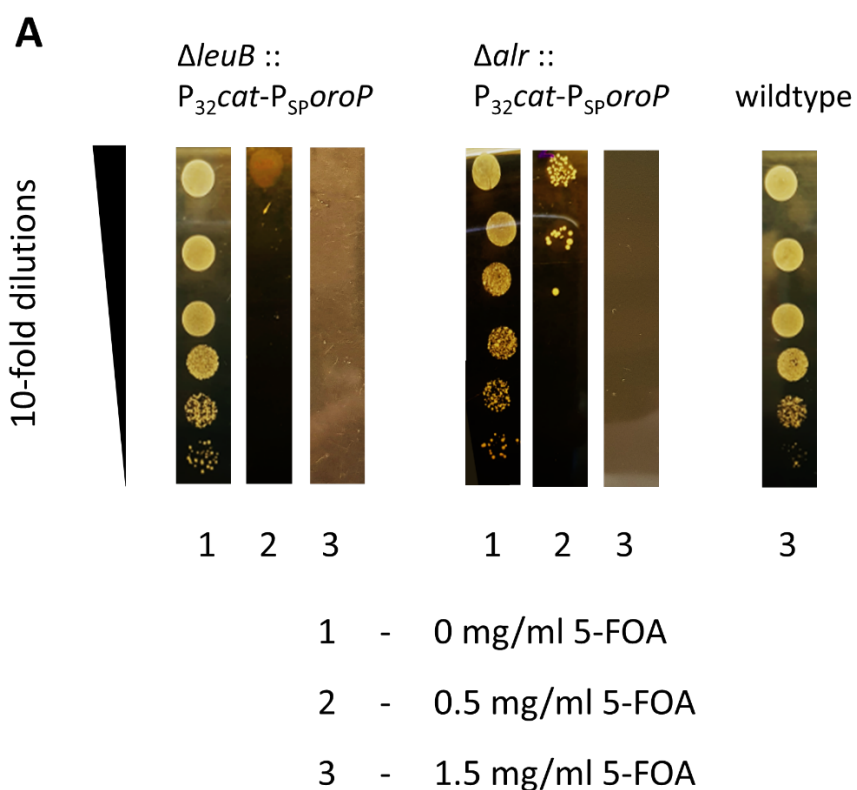
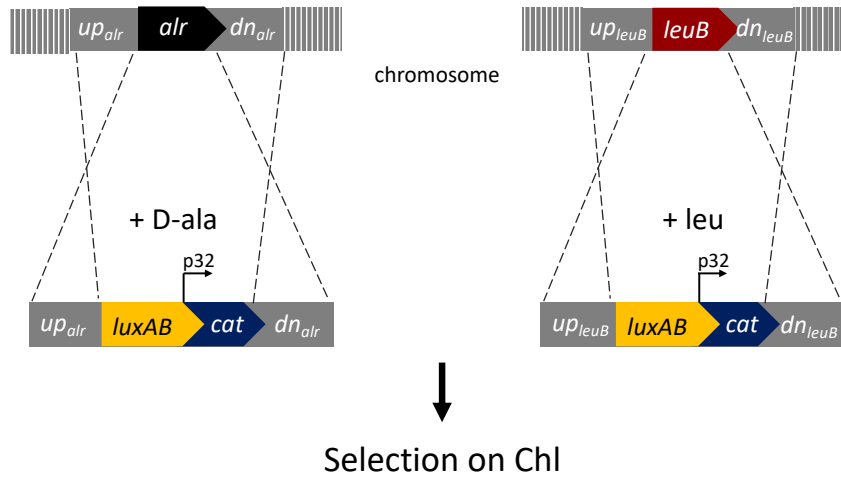


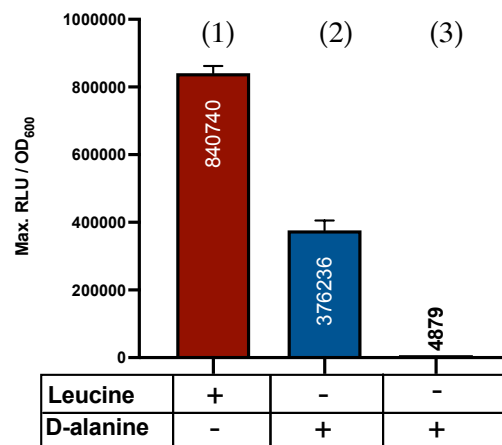
Figure 1. 3. 5-FOA sensitivity of *oroP* carrying strains is dependent on expression level (A) 5-FOA sensitivity of the mutants compared to the wildtype strain. 10-fold dilutions were spotted onto solid M17 media plates supplemented with different concentrations of 5-FOA.

Results Chapter 3.1

B



C



(1) - $\Delta leuB :: luxAB-p32cat$

(2) - $\Delta alr :: luxAB-p32cat$

(3) - $luxAB-p32cat$

Figure 1.3. (Continued) **(B)** Schematic representation of the cassette for replacing native *alr* and *leuB* genes with *luxAB* reporter for measurement

S. thermophilus as a suitable host for constructing large scarless variant libraries cloned and expressed from the chromosome of promoter activity, *luxAB* was cloned at the tRNAserine locus (with no promoter) as a negative control to demonstrate background expression level **(C)** The native *leuB* gene is expressed at a 2.2-fold higher level than the native *alr* gene *luxAB-P₃₂cat*. Luciferase activity expressed as a relative light unit (RLU) is reported as maximum RLU/OD₆₀₀. Monitored at 10 min intervals for 24 hours or less in a multi-well plate reader (Hidex Sense, Hidex, Turku, Finland) as previously described (Fontaine, Boutry, De Frahan, Delplace, Fremaux, et al. 2010).

***cat-oroP* based selection/counter-selection allows markerless replacement of genes at high transformation efficiencies**

Our aim was to evaluate *S. thermophilus* LMD-9 as a host for cloning and expressing large variant libraries. For this, we sought to estimate the maximum transformation efficiency we can achieve with natural transformation combined with the *cat-oroP* selection/counter-selection. We PCR amplified the wild-type *leuB* gene along with 1.5 kb flanking homology arms. We then used this as the donor DNA for natural transformation of the mutant $\Delta leuB :: P_{32-cat} P_{SP-oroP}$ strain **(Figure 1. 4 A)**.

We observe that the number of transformants increased with amount of donor DNA, plateauing around 10⁸ cfu/ml. 5-FOA resistant non-transformants (false positives) were observed in the control transformation mix without added donor DNA. For 2.5 x 10⁸ transformants, there is around 0.008% false positives. These false

Results Chapter 3.1

positives were observed to be predominantly originating due to deletions and nonsense mutations in the *oroP* marker.

The total volume of transformation mix can be easily scaled up to reach 10^9 transformants per natural transformation. Although, the bottleneck for achieving 10^{10} transformants is the prohibitively high amount of donor DNA required. Typical donor DNA yields for PCR amplification reactions are of the order of 1- 2 μg . This yield often decreases in subsequent Gibson/overlap PCR steps. Thus, a single 100 ml natural transformation reaction would require the equivalent of 5 ml PCR reactions. Since this was too resource intensive, we never estimated the upper limit of the transformants that could be achieved in a single transformation.

What is also interesting is that around 10% of the total viable cells can be transformed. This is in stark contrast to *E. coli* chemical and electro-competent cells where the maximal number of transformants have been reported to be around ~1% of total viable cells (often called “fraction competent” in previous research) (Bergmans, Van Die, & Hoekstra, 1981).

S. thermophilus as a suitable host for constructing large scarless variant libraries cloned and expressed from the chromosome

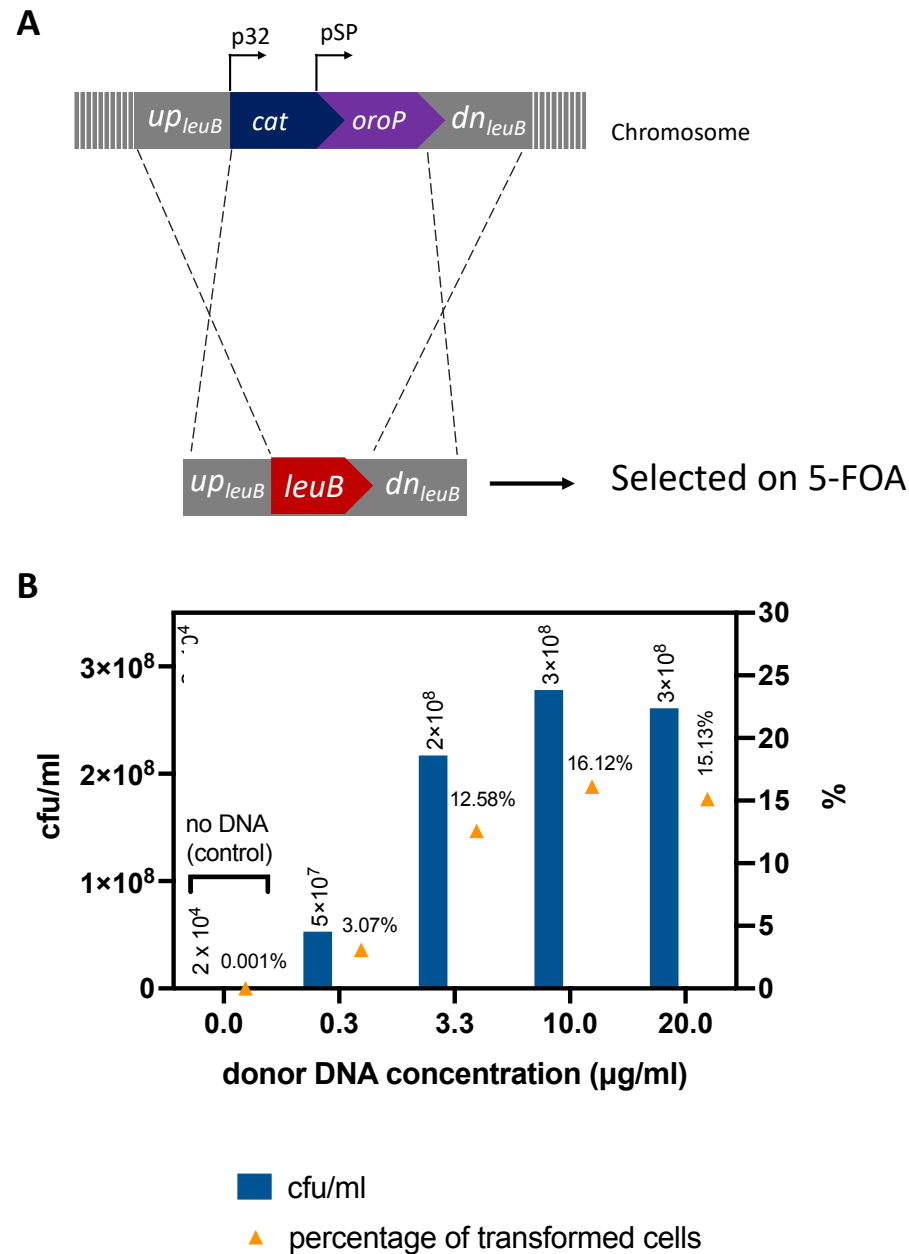


Figure 1. 4. Estimation of transformation efficiencies. (A) Schematic representation of the construct used for replacement of *cat-oroP* cassette with

Results Chapter 3.1

the wild-type *leuB* gene **(B)** cfu/ml and percentage of cells transformed. Cfu/ml was calculated by counting the number of colonies observed in serial dilutions spotted on growth media with 5-FOA supplementation. Percentage of cells transformed was calculated by counting the number of colonies observed in serial dilutions spotted on growth media after 24 hours of incubation on media supplemented with (transformed) and without 5-FOA (total viable cells), and subsequently calculating the percentage of transformed cells. False positives were calculated by observing the number of 5-FOA resistant colonies in the transformation without donor DNA.

Deletion of *lysA* gene from a the genome of cells encoding a library of *Alr* variants

We replaced the native *alr* gene with an error prone PCR library of variants. The $\Delta alr::P_{32-cat} P_{SP-oroP}$ strain was used as a recipient to insert an error-prone PCR library of variants (library preparation is discussed in more detail in Chapter 3.2). The error-prone PCR product was assembled with homologous arms by Gibson assembly (Galka et al. 2017), this was used as donor DNA for natural transformation. Serial dilutions of the transformation were spotted on plates supplemented with 5-FOA to reveal that we had a total of 3.29×10^8 transformants (**Figure 1. 5 A**). A total of 4.98×10^4 false positive 5-FOA resistant colonies were observed in the no DNA control. Transformants were spread on ten large 500 cm² square petri dishes to select for growth on media supplemented with 5-FOA.

S. thermophilus as a suitable host for constructing large scarless variant libraries cloned and expressed from the chromosome

Typically for variant libraries to be used for directed evolution or deep mutational scanning, library creation is the final step in the workflow. Any subsequent genetic background manipulation of the cells hosting the variant libraries would result in significant loss of variant diversity since transformative steps often result in comparatively few transformed cells and an abundance of non-transformed cells. Since natural transformation in *S. thermophilus* allows this combination of high transformation efficiency (10^9 cfu/ml) and high percentage of transformation (~10%). We decided to try and modify the genetic background of variants in the *alr* library in a single transformation. *lysA* is a gene involved in the lysine biosynthesis pathway in *S. thermophilus*. It encodes a Diaminopimelate decarboxylase involved in the decarboxylation of meso-diaminopimelate (meso-DAP) to L-lysine. As expected, a Δ *lysA* mutant of wild-type LMD-9 strain demonstrates complete L-lysine auxotrophy. To demonstrate that the modification of the genetic background of a library is possible, we decided to delete the *lysA* gene and thus modify the genetic background of the Alr library from a lysine prototroph to auxotroph.

Since each transformant in the Alr library is considered to be encoding a unique error-prone PCR variant of Alr, we started with 10 times more cfu than the initial diversity of the library to have probability of 95% library coverage. These cells were allowed to recover and undergo a few doublings in 200 ml skimmed milk media (SMM), and subsequently

Results Chapter 3.1

this was used as a starting point for natural transformation. A cassette comprising of up^{lysA} -P₃₂-*cat*- dn^{lysA} was assembled by Gibson assembly, the up and down homologous arms correspond to the genomic sequences 1.5 kb upstream and downstream to the *lysA* chromosomal locus. XIP and 40 µg of this donor DNA (0.2 µg/ml) was added for natural transformation (**Figure 1. 5 B**). Transformants were selected on media supplemented with chloramphenicol, D-alanine and L-lysine. Serial dilutions of the transformation spotted on growth media supplemented with chloramphenicol reveal a total of 2.0×10^9 transformants (**Figure 1. 5 C**). We pooled the transformants and subsequently plated dilutions on media supplemented without and with L-lysine as a control. We observe no colonies without L-lysine supplementation, demonstrating that all the cells are auxotrophic to lysine (**Figure 1. 5 D**).

S. thermophilus as a suitable host for constructing large scarless variant libraries cloned and expressed from the chromosome

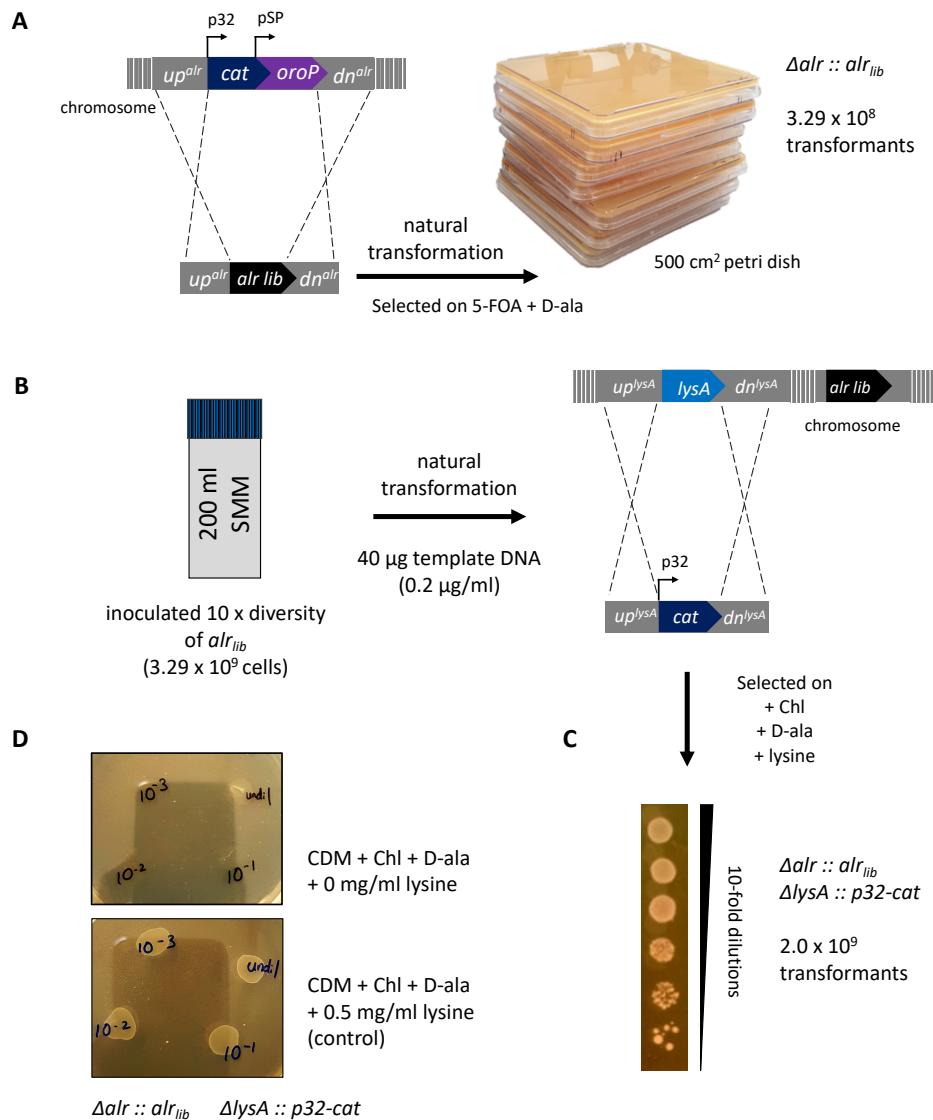


Figure 1. 5. Changing the genetic background of a strain after generation of a library. (A) Schematic representation of the strategy used for generation of a library of *alr* variants. $\Delta alr :: P_{32} cat-P_{SP} oroP$ strain was

Results Chapter 3.1

used as a recipient to insert a library of ep-PCR *alr* variants by natural transformation. 3.2×10^8 transformants were selected on growth media supplemented with D-alanine and 1.5mg/ml 5-FOA. **(B)** Strategy used for deletion of *lysA* gene from the cells encoding the library of *alr* variants. Ten times the initial number of transformants of the *alr* library 3.29×10^9 cells were inoculated into 200ml of skimmed milk media (SMM), 40 μ g of donor DNA cassette consisting of $\Delta lysA :: P_{32} cat$ was added to replace the *lysA* gene **(C)** Phenotypic characterization of the $\Delta lysA$ mutant library. 2×10^9 transformants were selected on solid growth media supplemented with D-alanine, lysine and chloramphenicol. **(D)** Dilutions of transformants plated on media with and without lysine supplementation demonstrate development of lysine auxotrophy in the transformed cells.

3.1.4. Discussion

In this work, we have shown that it is possible to create scarless mutants in the genome of *S. thermophilus* via a simple 2-step procedure. We demonstrate how this method can be employed to generate large chromosomally encoded libraries of genes cloned at their native locus. Later, we demonstrate changing the genetic background of the variants of a library from lysine prototroph to auxotroph. We could envision a scenario where the ability to change the genetic background of a library could come in use. We could select for enzyme activities complementing an essential function by cloning a variant library in the genome and subsequently deleting the essential gene under question for selection of

S. thermophilus as a suitable host for constructing large scarless variant libraries cloned and expressed from the chromosome positive variants. Similarly, transforming a library with a gene encoding a toxic protein may be an interesting way to discover inhibitory compounds.

S. thermophilus is a U.S. FDA approved “*Generally Recognised As Safe*” (GRAS) status bacteria, thus it needs fewer regulatory approvals for industrial use. A small genome (1.86 Mb), well-studied physiology (Hols, Hancy, Fontaine, Grossiord, Prozzi, Leblond-Bourget, Decaris, Bolotin, Delorme, Ehrlich, et al. 2005) and efficient natural transformation offers versatile uses in directed evolution and synthetic biology. For example, to be used for metabolic engineering, since large (15 kb) operons can be cloned easily (Dandoy et al. 2011), in continuous evolution systems, since bacteria can be easily transformed directly in growth media without electroporation or heat-shock, additionally it could also be used as a model to evolve reduced genomes in bacteria or to study whole genome transplantation. As a vector for directed evolution of proteins, the ability to anchor proteins on the surface (Lecomte et al. 2016; Mu et al. 2020) permits use of droplet microfluidics/ fluorescence-activated cell sorting for screening improved variants. The disadvantages of *S. thermophilus* include, the less developed set of genetic tools (inducible promoters, selection markers etc.) as compared to standard model organisms like *E. coli*. requirement of maintaining anaerobic conditions.

3.1.5. Future work

To further optimize the counter-selection system the *oroP* gene could be used as both a selection and counter selection marker (Defoor, Kryger,

Results Chapter 3.1

and Martinussen 2007). To do this in *S. thermophilus*, it would involve engineering pyrimidine auxotroph mutants by tightly controlling the expression of a gene in the orotate biosynthesis pathway by means of an inducible promoter. In uninduced conditions, this engineered strain would be dependent on the import of orotate by OroP for biosynthesis of uridine monophosphate (UMP). Subsequently, the gene can be induced and *oroP* can be counter-selected.

In *S. thermophilus*, orotate phosphoribosyl transferase (STER_RS04840) and orotidine-5'-phosphate decarboxylase (STER_RS04835) are the enzymes involved in this pathway. An inducible promoter could be introduced by natural transformation, and screening for uninduced-induced pyrimidine auxotroph-prototrophs. Thus allowing use as both a positive and negative selection marker. This bypasses the requirement for the *cat* gene as a positive selection marker, and would reduce false positives generated via accumulation of mutations in *oroP*.

3.2. Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *Streptococcus thermophilus*

Jason Baby Chirakadavil, Steve Alarcia, Salwa Ben Abbou, Pascal Hols, Patrice Soumillion

Lab of Biochemistry and Genetics of Microorganisms (BGM), Louvain Institute of Biomolecular Science and Technology (LIBST), Université catholique de Louvain, Place Croix du Sud, 4-5 box L7.07.06, 1348 Louvain-La-Neuve, Belgium.

Author Contributions:

J.B.C., P.S. and P.H. conceived the idea and designed the experiments; J.B.C. performed the experiments, analysed the data and wrote the manuscript; P.S. and P.H. reviewed and edited the manuscript; J.B.C. and S.A. performed the bioinformatic analyses; S.B.A. contributed to the Alr library preparation.

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

3.2.1. Abstract

The study of how enzymes evolve to adapt to selection pressure has been a central theme in molecular evolution. Using high-throughput next generation sequencing (NGS), we can quantitatively study the selection and purge of evolving genotypes and subsequently use this information to map the effects of mutations of all enzyme residues on bacterial fitness. These experiments are often performed using recombinant expression from multi-copy plasmids that can generate biases due to much higher protein concentration compared to natural physiological conditions. This can lead to either pervasive mutational tolerance or artificial fitness costs. Here, we took advantage of the natural competence of *S. thermophilus* to develop a new experimental system to clone gene libraries directly from the native chromosomal locus with extremely high efficiency (upto 10^9 transformants). A library of bacteria expressing alanine racemase (Alr) variants was created and grown for 8 days while allowing fitness effects to manifest as on the growth rate of the bacteria. The fate of 1712 substitutions were tracked in conditions with or without selection pressure for the native Alr activity.

Under selective pressure for maintaining alanine racemase activity (growth without D-alanine), we observe that 43.57% of substitutions were “highly deleterious”, as expected these included premature stop codons, active site residues and mainly buried residues. In general, biophysical

Results Chapter 3.2

constraints (exposed hydrophobic residues) or functional constraints (proximity to active site, structurally important position) could explain most of the purged mutations with a few exceptions. We also observe 1.52% of “slightly beneficial” mutations. These include D170E and R288K situated at the middle layer of the active site entryway, and preliminary results indicate that substitutions could reduce access to the active site, possibly modulating the native and promiscuous activities towards an improved fitness.

In drift without selection pressure for alanine racemase activity, we observe predominantly neutral fitness effects of 66.65% substitutions, 0% highly deleterious substitutions and 2.69% “beneficial” selections. These beneficial substitutions also included the substitutions D170E and R288K. Additionally, Y263C in the catalytic centre was also selected and may indicate the emergence of a beneficial promiscuous activity.

Our method, mimicking the constraints of natural molecular evolution, provides a tool to study distributions of fitness effect of mutations. This could open doors to study fundamental aspects of protein evolution, opening new research strategies.

3.2.2. Introduction

There is a lack of experimental data which accurately maps the effect of mutations on organismal fitness. Studies on this topic either deal with this question computationally by mining existing sequence

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

information through databases (Kulshreshtha et al. 2016; Mazurenko 2020) or have experimental systems which can create artefactual selection or purge of mutations. Many previous studies have highlighted contradictory results regarding the tolerance of the native enzyme activity to mutations. Bershtein et al. show that drift of TEM-1 beta-lactamase under selection for antibiotic resistance yielded greater genetic diversity encoding the native genotype and thus increases the propensity for selection of new activities (Bershtein, Goldin, and Tawfik 2008). Subsequently they have shown that these consensus genotypes which are mutationally robust can evolve into new enzyme activities (Aharoni et al. 2005; Khersonsky and Tawfik 2010), leading them to propose the model linking robustness (larger number of genotypes encoding native enzyme activity), which allows the protein to occupy the 'neutral network' (Bornberg-Bauer and Chan 1999), and increase evolution of promiscuous activities (Tawfik 2010). However, in nature, selective pressure for beta-lactamase is not always present and fitness costs associated with expression of beta-lactamase mutants have also been reported (Mehlhoff et al. 2020). It is therefore unclear how diverse a natural genotype can be. More recent studies have also challenged the overall idea of a mutationally robust native enzyme activity. Studies for evolutionary transitions in enzyme function: (1) from phosphotriesterase to arylesterase, and (2) from atrazine chlorohydrolase to melamine deaminase have shown that 67% of mutations in the native enzyme were

Results Chapter 3.2

deleterious, furthermore they propose that increases in promiscuous activities are independent to any change in native activity and often involve a weak trade-off just as a by-product of stochastic epistatic effects of selection for new activity (Kaltenbach et al. 2016). On the other hand, Copley's team demonstrated a strong native activity trade-off concurrently along with increase in promiscuous activity with a single Glu to Ala mutation in gamma-glutamyl phosphate reductase, which created physiologically relevant promiscuous activity for multiple orthologs tested (Khanal et al. 2015). Hall et al. have also demonstrated a strong trade-off under selection between native and promiscuous activity mainly driven by decrease in affinity for native substrate. Selection for improvement in new activity was possible only after complete loss of native activity, leading them to propose that relatively few mutations can allow for a complete change in enzyme specificity (Hall et al., 2020).

Two consecutive studies on Hsp90 chaperone from the Daniel Bolon's team provide a good example of the importance of controlling expression levels in fitness mappings, initially mutational fitness mapping revealed most mutations being fitness-neutral (R. T. Hietpas, Jensen, and Bolon 2011). However, they later mapped the fitness effect (growth rate) to the expression level of Hsp90, revealing a non-linear elasticity function such that expression level could be reduced upto ~79% without a concomitant measurable effect on cell growth. Fitness effects which were previously "hidden" to growth revealed themselves at reduced expression

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

levels (Jiang et al. 2013). This has also been seen with the estimation of the fitness effect of mutations in a DNA methyltransferase, where mutations were purged at reduced expression levels resulting in ~66% of mutations being deemed deleterious (Rockah-Shmuel, Tóth-Petróczy, and Tawfik 2015). Even in low expression conditions, the use of plasmid expression systems leads to high protein concentration compared to physiological conditions. Plasmids also generate heterogeneous enzyme expression levels between variants due to copy number variation. Ryan et al. demonstrate that these problems could be resolved by moving to chromosomal expression systems. Tawfik's team established another elegant experimental system in *Bacillus subtilis*, where the glutamate dehydrogenase (GudB) was randomised at 10 individual positions with an NNS codon (Noda-García et al. 2019). These mutations were integrated into the chromosomal *gudB* gene and the fitness of these mutants were estimated under competition in different environments. It was observed that most mutations were deleterious, while the wild-type *gudB* sequence was best fitted when seen across multiple conditions, they suggest that laboratory conditions typically underestimate the fraction of mutations that are deleterious in nature.

Numerous computational predictors of protein mutational effects have been developed. These methods have been trained on a combination of large scale mutational data sets, clinical data and presumed neutrality of substitutions from closely related orthologous

Results Chapter 3.2

proteins (Wei et al. 2010). A few of these methods like SIFT and PROVEAN use phylogenetic data to predict the deleterious effect of mutations. (Choi et al. 2012; P. C. Ng and Henikoff 2001). These methods have been tested and validated using different mutational datasets. In our work, we will test our dataset of experimental values of Alr mutations and their effect on the enzyme function. Alr is essential for bacterial growth, the mutant Alr has to fold, dimerise, bind to the cofactor and catalyse the reaction for growth. Since the experiment is a functional test of protein activity in a biologically relevant context, mutations which are deemed deleterious or neutral in experimental conditions can be compared to computationally predicted values to evaluate the accuracy of the predictors.

Alr are predominantly only present in bacteria and some fungi, thus it is an interesting target for antibacterial drug design. Notably, D-cycloserine (DCS) is an Alr inhibitor and a broad spectrum antibiotic currently used to treat infection by multi-drug resistant *Mycobacterium tuberculosis* infections (Caminero et al. 2010). Recently, DCS was also approved by the Food and Drug administration for the treatment of refractory enteric infections by *Clostridium difficile*. DCS was initially considered as an irreversible covalent inhibitor of alanine racemase (Fenn et al. 2003), but recently it has been shown to be a slow, reversible covalent inhibitor instead (de Chiara et al. 2020). In *M. tuberculosis*, DCS also inhibits D-Ala:D-Ala ligase (Ddl) yet most clinical isolates showing DCS

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

resistance implicate mutations in the promoter and coding sequence of Alr (Evangelopoulos et al. 2019). A number of these Alr mutants have been purified and their effect on DCS inhibition and racemase activity have been estimated (Nakatani et al. 2017). We know that these mutations reduce racemization activity of Alr but it is not clear whether this reduction in activity is strong enough to be subject to negative selection, since many enzymes are expressed at levels higher than required for physiological function (Newton, Arcus, and Patrick 2015). Furthermore, even overexpression of Alr can confer DCS resistance in *Mycobacterium smegmatis* (Cáceres et al. 1997), emphasizing the need to estimate the effects of the mutants at native expression levels. We lack a comprehensive study of the implications of Alr mutations to cellular fitness in the absence of DCS driven selection pressure. Secondly, rational inhibitor design strategies could benefit from a map of conserved regions to target Alr, but merely relying on evolutionary conservation values does not allow us to infer the local constraints of available residues for a particular position in Alr evolving under natural expression conditions. This study also aims to contribute to answering these questions.

In the previous chapter, we have presented a new method for building libraries of genes cloned at the chromosome in *S. thermophilus* (Section 3.1). In this work, we employ this technique to create a highly diverse library of Alanine racemase (Alr) mutants cloned at its native chromosomal locus. We then grew this library for 8 days in two conditions

Results Chapter 3.2

- (1) with D-alanine supplementation (2) without D-alanine supplementation, to study the selection and purge of variants in both conditions.

3.2.3. Results

Availability of D-alanine is linked to organismal growth rate, allowing us to create a genotype-phenotype linkage which directly correlates with activity of AlrSt

We initially estimated the effect of deletion of the gene *alr* on growth. As previously described (Section 3.1), using the *cat-oroP* two-step cloning strategy we replaced the *alr* gene by (1) a cassette consisting of the up^{alr} and down^{alr} flanking regions creating the Δalr strain, and (2) a cassette consisting of up^{alr} arm, *alr* with a strep tag (denoted as *alr(strep)* for the rest of this work), down^{alr} arm creating the $\Delta alr:: alr(strep)$ strain (Figure 2. 1 A). The transformants were selected by growing them on media complemented with 5-fluoroorotic acid (5-FOA). The transforming DNA was prepared by PCR and Gibson assembly, and natural competence was activated in *S. thermophilus* LMD-9 by addition of ComS peptide as described in materials and methods (Chapter 6). These strains were subject to growth with and without addition of external D-alanine in the growth medium. We observe deletion of *alr* renders the bacterium auxotrophic for D-alanine. Furthermore, the growth rate is depending on the amount of available D-alanine, with growth rate plateauing at 0.125 mg/ml (Figure 2. 1 C). The *alr(strep)* gene fully restores the D-alanine

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus* prototrophy (Figure 2. 1 B). Addition of D-alanine does not contribute to any detectable benefit or cost to the growth rate of *alr*(strep). Unless explicitly mentioned otherwise, whenever D-alanine was added to the medium to complement growth, it was added at a concentration of 0.5 mg/ml.

Results Chapter 3.2

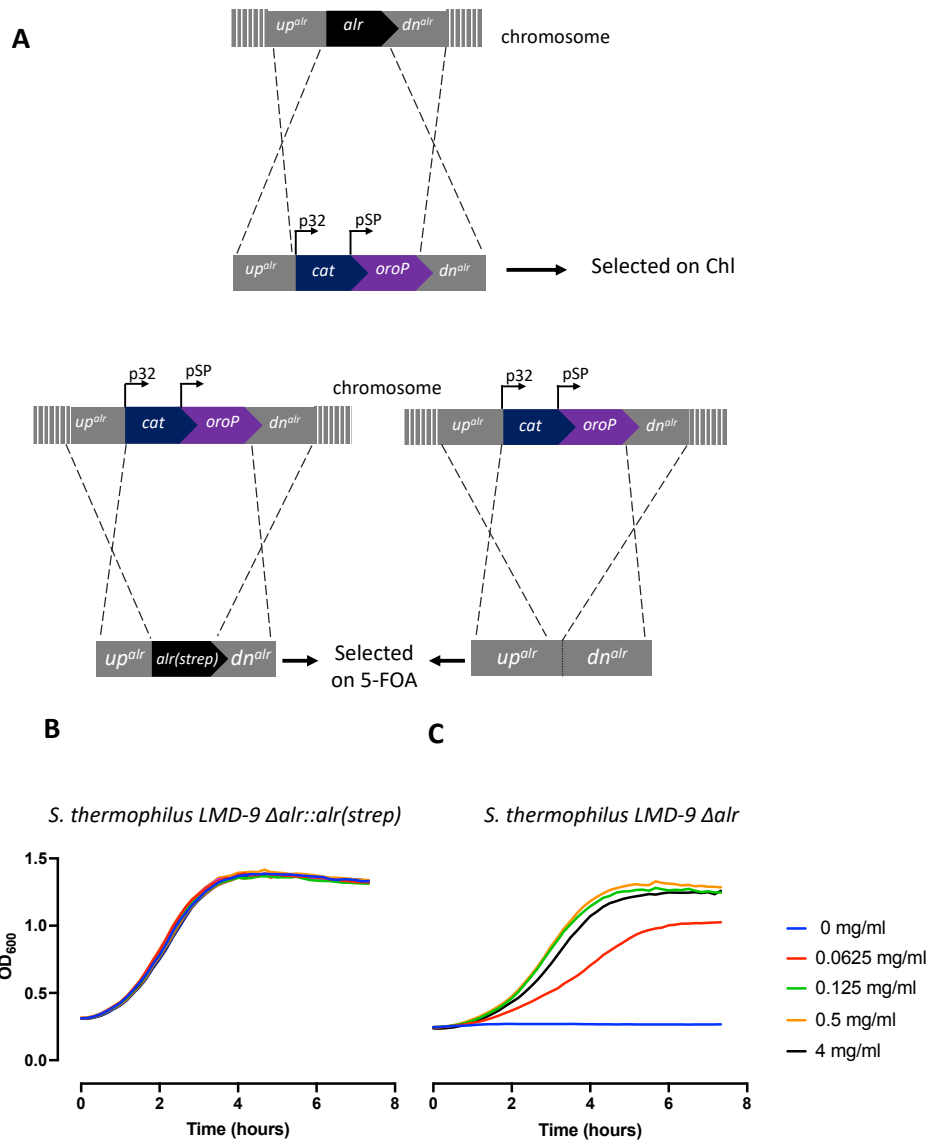


Figure 2. 1 **(A)** Scheme of the construction of the strains **(B)** Replacing the wild-type chromosomal *alr* with *alr(strep)* completely restores D-alanine auxotrophy. Deletion of *alr* results in D-alanine auxotrophy. **(C)** Addition of D-alanine complements growth in a Δalr strain. Growth was measured

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus* in triplicates at 10 minutes intervals on a Tecan Infinite 200 PRO series (Tecan, Männedorf, Switzerland). Mean values calculated from three independent replicates are represented here.

Alr is expressed constitutively in *S. thermophilus*

We know that Alr is required by the *S. thermophilus* LMD-9 for growth. However, it is yet to be identified whether expression of Alr would be constitutive or induced by presence or absence of D-Alanine or L-Alanine. Analysis of the region directly upstream of *alr* reveals a single Shine-Dalgarno sequence at position 1592400 (AGGAGA) 6 nucleotides upstream of *alr* ORF. Upstream of *alr*, we found two genes (STER_RS08320 and STER_RS08325) before the first intergenic region (97 bp, Figure 2.2.A) in which we identified partially conserved -10 (Pribnow box) at position 1593830 (TATTT), and -35 box at position 1593855 (TTGGGA) (**Figure 2. 2 A**). We annotated this region as a putative *alr* promoter (P_{alr}) and selected it for further analysis.

We created two *luxAB* luminescence reporter strains to monitor transcriptional activity of the *alr* promoter. (1) In the first one, *luxAB* gene is transcriptionally fused to P_{alr} at an ectopic locus. Transcriptional fusion P_{alr} -*luxAB* was inserted into the STER_RS07710 locus (annotated as *tRNA_{serine}*), previously reported to be permissive to insertions (Mignolet et al. 2018). This reporter strain was grown in chemically defined media (CDM). D- or L-alanine were added externally to test their effects on the

Results Chapter 3.2

expression of P_{alr} -*luxAB*. (2) For the second reporter strain, $\Delta alr::luxAB$, the native *alr* gene was replaced by the *luxAB* gene. This strain is auxotrophic to D-Alanine and allow to evaluate transcriptional activity only when L- and D-Alanine are present in the medium.

P_{alr} seems to be unaffected by addition of L-alanine or D-alanine. We observe a similar level of luminescence activity in presence of D-alanine, for *luxAB* cloned at native and ectopic *tRNA_{serine}* locus (**Figure 2. 2 B**). However, simultaneously removing both L-alanine and D-alanine reduces the expression levels by 26 percent (**Figure 2. 2 B**), this might be explained by the added metabolic pressure for synthesizing L-alanine through pyruvate by the alanine dehydrogenase (STER_RS07855) or possibly via aspartate catabolism through an undefined aminotransferase (Hols, Hancy, Fontaine, Grossiord, Prozzi, Leblond-Bourget, Decaris, Bolotin, Delorme, Dusko Ehrlich, et al. 2005). This suggests a possible additional level of expression control.

When *luxAB* was inserted in place of *alr*, we also observed a very similar luminescence, suggesting that the ectopic site does not create biases in transcriptional activity. Overall, our results indicate that the expression seems to be constitutive.

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

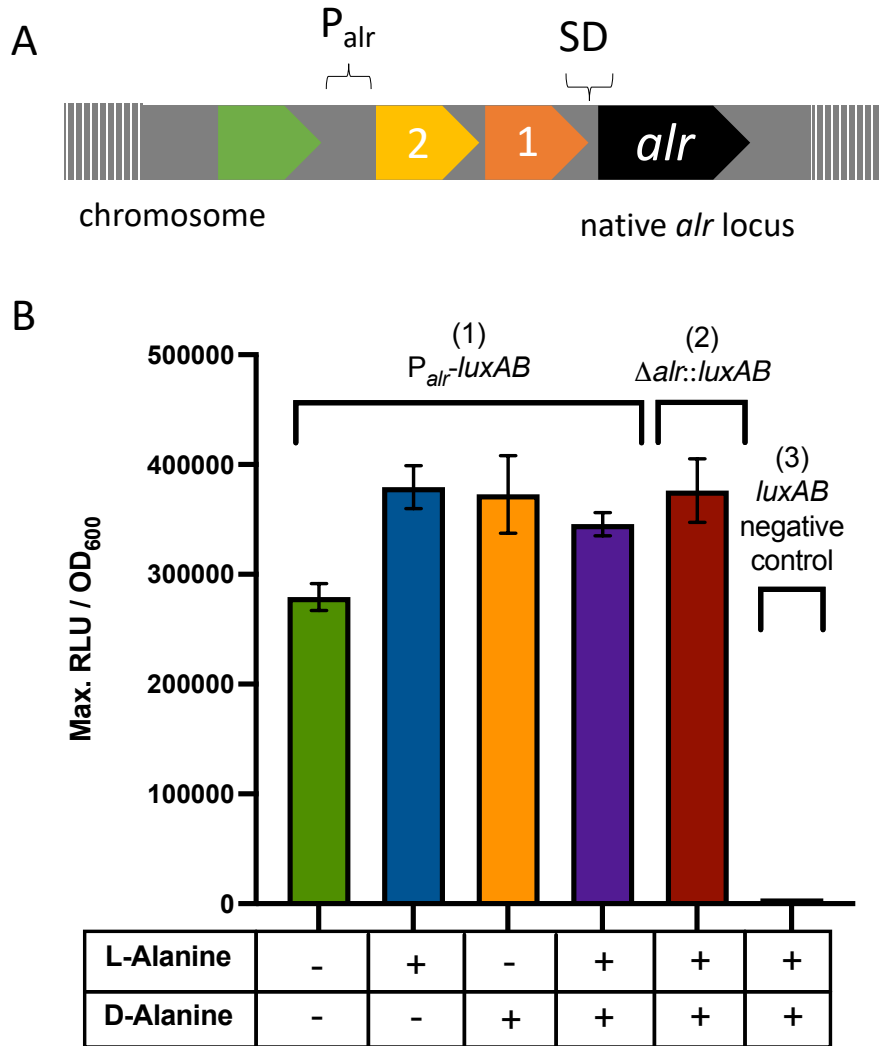


Figure 2. 2 (A) Organization of the *alr* promoter at the native *alr* locus. SD denotes the position of the Shine-Dalgarno sequence. Genes labelled as 1 and 2 in the figure stand for STER_RS08320 (annotated as a holo-ACP synthase based on homology) and STER_RS08325 (annotated as a 3-deoxy-7-phosphoheptulonate synthase based on homology) respectively

Results Chapter 3.2

(B) Expression studies conducted with (1) transcriptional fusion of P_{alr} , and (2) replacement of the native *alr* gene by the *luxAB* reporter, shows that expression of *Alr* is not affected by the presence of L- or D-alanine and it is probably constitutively active. (3) *luxAB* was cloned at the *tRNAserine* locus (with no promoter) as a negative control to demonstrate the background expression level. Growth (OD_{600}) and luciferase activity (expressed as a relative light unit; RLU) were monitored at 10 min intervals for 24 hours or less in a multi-well plate reader (Hidex Sense, Hidex, Turku, Finland) as previously described (Fontaine, Boutry, De Frahan, Delplace, Fremaux, et al. 2010).

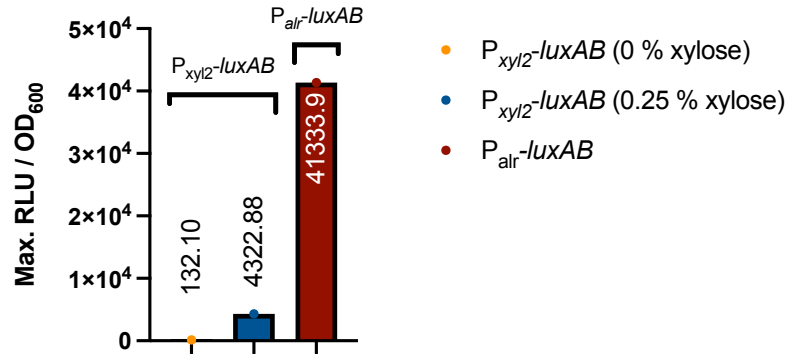
***Alr* expression can be reduced without significant effect on growth**

Previous studies have shown that a non-linear relationship between growth rate and protein expression levels exists for certain proteins. Newton et al. describe many examples where huge reductions in enzyme activity do not have a concurrent reduction in growth rate (Newton, Arcus, and Patrick 2015). Similarly, it has been shown that the expression level of the chaperone *Hsp90* has to be drastically reduced to start developing an observable growth defect (Jiang et al. 2013). We decided to estimate the extent of this relation for *Alr*. First, we cloned the *luxAB* reporter gene under a xylose inducible promoter *Xyl-S1* (denoted as P_{xyl2} in this study) (Xie, Qi, and Merritt 2013), creating the transcriptional reporter P_{xyl2} -*luxAB* strain. Next, we compared the induction level of P_{xyl2}

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus* under basal (0 % xylose) and maximum (0.25% xylose) induction levels, with the previously described *P_{alr}-luxAB* strain. We see that under maximum induction levels for *P_{xyI2}, P_{alr}* has 9.5 times higher expression level, whereas under basal expression level for *P_{xyI2}, P_{alr}* has roughly 312 times higher expression level (**Figure 2. 3 A**). Next, we deleted the native *alr* gene and cloned *alr* under the control of *P_{xyI2}* creating the *P_{xyI2} -alr* strain. We induced Alr at basal and high induction conditions to estimate the relation between Alr expression and growth rate in the absence of D-Alanine. Surprisingly, both conditions displayed wild-type like fitness, with no significant effect of growth even in basal induction conditions. We were unable to determine the lower limit of expression strength at which we would start to see a detrimental effect on the bacterial growth rate. This suggests that intracellular concentration of wild-type Alr can be considerably reduced before we start observing detrimental effects on cell growth.

Results Chapter 3.2

A



B

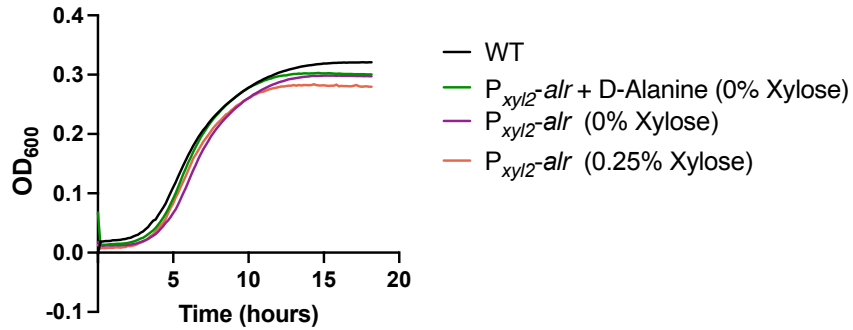


Figure 2. 3. (A) Expression level of P_{alr} compared to different levels of induction of P_{xyl2} . **(B)** Reduced expression of *alr* does not directly result in a concomitant reduction in bacterial growth rate suggesting that the activity of wild-type Alr is orders of magnitude higher than what is required for cellular metabolism. Growth was measured in triplicates at 10 minutes intervals on a Tecan Infinite 200 PRO series (Tecan, Männedorf, Switzerland). Mean values calculated from three independent replicates are represented here.

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

Homology Modelling of AlrSt

The structure of AlrSt was modelled using SWISS-MODEL (Bordoli and Schwede 2012; Waterhouse et al. 2018). As shown in **Figure 2. 4**, the models superimpose well with the experimental structure of Alr^{Sp} (pdb: 3S46) with whom AlrSt shares 61% sequence identity without any indel in the pairwise alignment.

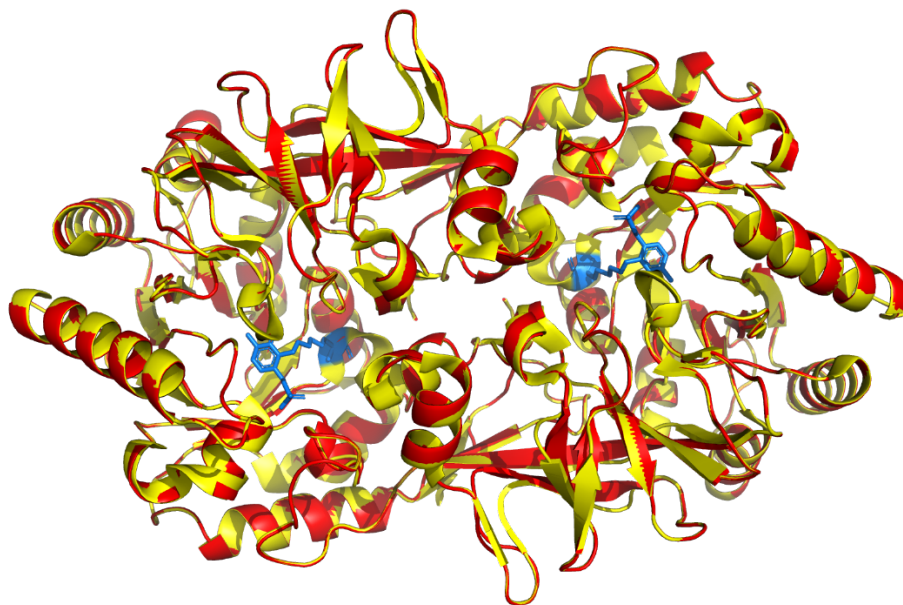


Figure 2. 4. Predicted model of AlrSt. The proteins are represented as cartoons, with AlrSt in red, Alr^{Sp} (pdb:: 3S46) in yellow. The active site LLP is shown in blue. The root mean square deviation (RMSD) between the 4173 atoms was 0.085, as calculated by PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC).

Generation of a library of Alr mutants expressed from its native chromosomal locus

We generated a library of Alr mutants by error-prone PCR. Primers were designed and used to PCR amplify the *alr(strep)* gene using the strain $\Delta alr::alr(strep)$ as template. This PCR product was used as the template for the error prone PCR. The quantity of starting template and number of cycles were optimized to obtain 0 - 4.5 mutations/kb. Regions flanking the *alr* gene, up^{alr} and down^{alr} were PCR amplified, and finally assembled together with the error-prone PCR product by Gibson reaction providing the up^{alr}-*alr^{lib}*-down^{alr} cassette. The error-prone PCR product encoding the Alr mutant library would be denoted as *alr^{lib}* for the rest of this study. The *S. thermophilus* LMD-9 strain $\Delta alr::P_{32-cat}-P_{sp}-oroP$ which we previously described (Section 3.1) was utilized as a recipient strain for inserting the *alr^{lib}*. Natural competence was induced in this strain with XIP, and the up^{alr}-*alr^{lib}*-down^{alr} cassette was added as donor DNA for natural transformation (**Figure 2. 5 A**). Transformants were plated on large petri dishes (500cm²) containing M17 media + 5-FOA (1.5 mg/ml). Serial dilutions were prepared and spotted onto M17 plates with and without 5-FOA to estimate the number of transformants and the total viable cells respectively. Nine transformations were carried out in parallel generating a total of 3.29×10^8 transformants (N_0) among 3.65×10^9 total viable cells. Owing to the high natural competence of LMD-9 strain, we therefore achieved transformation of around 9% of the total viable cells.

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

For deep sequencing of the initial library, colonies were scraped from the large plates, lysed and the genome was purified and extracted. 6 µg of the extracted genome was used as an input for a 100 µl PCR with primers amplifying the *Alr* locus in the genome. The reason for adding such a high amount of template was to prevent bottlenecking of the high diversity of the initial library (6 µg corresponds to 3.15×10^9 copies of the genome of *S. thermophilus* (1.8 Mb) $\sim 10 \times N_0$). The PCR amplicons were sequenced by Ion Torrent sequencing. The *alr* gene with 366 codons randomised with 4 possible nucleotides at each 1098 positions, we have 4392 possibilities at the nucleotide level. Additionally, with 20 amino acid alternatives we would have 7320 possible amino acid mutations. Deep sequencing of the library revealed that at the nucleotide level, there are 4041 out of 4392 (92%) mutations with at least 10 counts in the sequencing (**Figure 2. 5 B**). At the amino acid level, we see 1712 missense and nonsense mutations in the library out of 7320 possibilities, with each residue of *Alr* on average mutating to 3 other residues (**Table 2. 1**).

We had calibrated epPCR conditions to achieve a low rate of mutation (0-4 mutations) for each transformant. Yet, since the actual distribution of mutations is expected to follow a Poisson curve. Eighteen colonies picked at random were Sanger sequenced to estimate the average distribution of mutations. 77.7% (14/18) of sequenced colonies encoded the full gene without insertions or deletions. For these, the number of missense + nonsense mutations per gene follows a Poisson

Results Chapter 3.2

distribution with a mean of 4.1. Deep sequencing of the d_0 library shows that we have a mean of 2.9 mutations/gene (**Table 2. 1**). Estimation of the distribution of the percent of single mutants in the library (**Figure 2. 5 C**), shows that 16% of our library would consist of variants encoding a single missense/nonsense mutation. This corresponds to 5.2×10^7 out of 3.28×10^8 total variants.

Table 2. 1. Parameters of the NGS data analysis of the initial library

Parameter	Quantity	Percent of total reads
Total number of reads	435179 reads	
Average length of reads (base pairs)	124 bp	
Reads shorter than 65 nucleotides (removed)	12442 reads	2.86%
Low quality reads (Q-value _{avg} < 25) (removed)	17625 reads	4.05%
Reads mismatching the template (removed)	76987 reads	17.69%
Reads aligned to template (retained)	328125 reads	75.40%
Missense + nonsense mutation load/gene	2.9 mutations/gene	

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

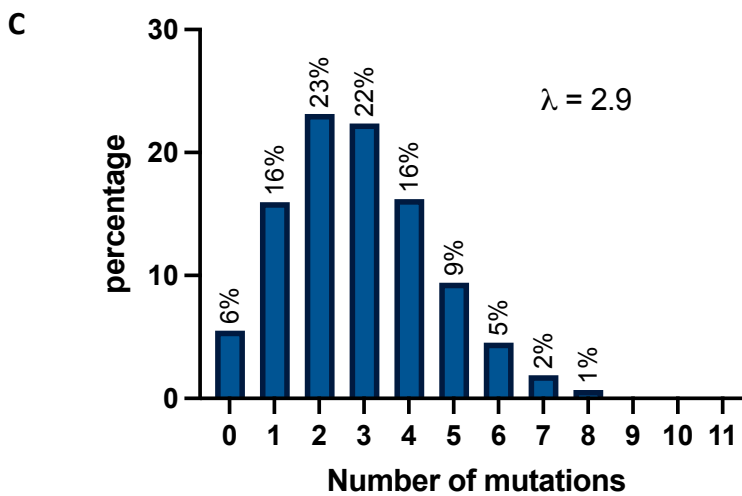
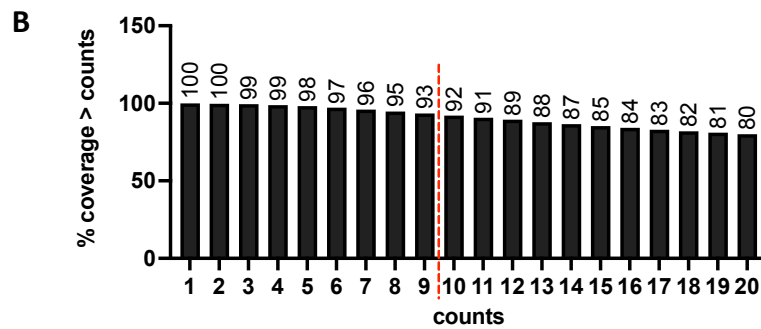
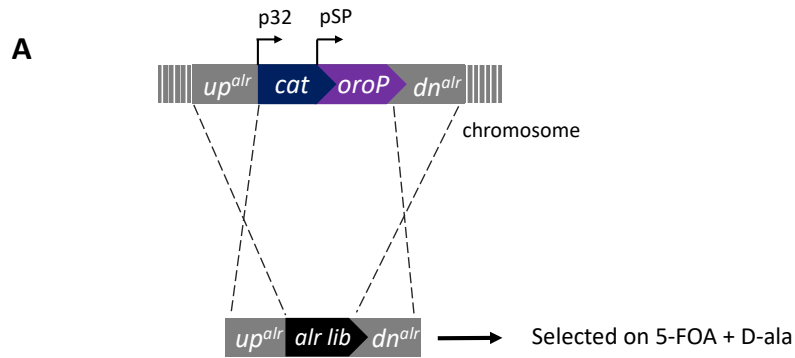


Figure 2. 5. Construction and preliminary characterization of the library of *alr* mutants **(A)** Mutants were selected by a scarless counter-selection strategy by selecting the cells against loss of the *oroP* gene and 5-FOA sensitivity. **(B)** Percentage of mutants in the library counted for at least the indicated number of times. (red line). **(C)** Poisson distribution of the number of mutations carried per variant. The value of rate parameter, λ , was calculated from the counts from the NGS sequencing of the initial library, and is displayed on the graph.

Mutations with low frequency reads in initial library (d_0)

Next, we decided to explore the mutations very low in the initial (d_0) library. The idea here is to look for potential mutations generating variants of Alr that may be toxic to the cell, for instance, by catalysing promiscuous reactions that would impede the metabolic equilibrium of the cell. Since the *alr*_{lib} is under constitutive induction, we reasoned that these toxic mutations would be rapidly purged out after the construction of the library. We searched the library at d_0 at the nucleotide level for low frequency mutations ($N_m^0 < 3$). Twenty-six sites in Alr have low frequency mutations. There are two possible explanations why the frequencies of these mutations are low: either they encode an amino acid creating a toxic mutant and thus get subject to purifying selection, or they are rare substitutions originating from the bias inherent to mutazyme II polymerase. To distinguish between these two possibilities, we looked at the frequency of the different nucleotide substitutions reported for

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

mutazyme II polymerase (**Table 2. 2**). Most missing substitutions originate from the frequency bias of the polymerase, moreover presence of alternate codons for some of the missing substitutions rule out the hypothesis of toxicity of mutants (Table 2. 3).

Table 2. 2 Substitution frequencies for Mutazyme II (Manufacturer's manual). Colours refer to the missing transitions referred in (Table 2. 3).

Transitions	Frequency
<i>A → G, T → C</i>	17.5 %
<i>G → A, C → T</i>	25.5 %
Transversions	
<i>A → T, T → A</i>	28.5 %
<i>A → C, T → G</i>	4.7 %
<i>G → C, C → G</i>	4.1 %
<i>G → T, C → A</i>	14.1 %

Table 2. 3 Low frequency mutations predominantly originate due to rare substitutions of the polymerase.

Results Chapter 3.2

Amino acid	Missing	Original Codon	Missing codon	Alternate codon present in library?
Ile 82	T → G	ATA	AGA	-
Ile 84	A → C	ATT	CTT	-
Val 93	T → G	GTT	GTG	-
Gln 115	C → G	CAG	GAG	-
Gly 118	G → C	GGG	CGG	Yes
Asp 120	A → C	GAT	GCT	-
His 127	C → G	CAT	GAT	-
Asp 131	A → C	GAT	GCT	-
Lys 143	G → C	AAG	AAC	Yes
Asp 144	G → C	GAC	CAC	-
Ile 149	A → C	ATT	CTT	-
Ile 149	T → C	ATT	ACT	Yes
Gly 161	T → G	GGT	GGG	-
Ile 162	A → C	ATT	CTT	-
Glu 171	G → C	GAA	CAA	-
Phe 177	T → G	TTT	GTT	-
Asp 189	A → C	GAT	GCT	-
Pro 195	T → G	CCT	CCG	Silent
Ile 207	A → C	ATT	CTT	-

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

Asn 215	A → C	AAT	CAT	-
Val 217	A → C	GTA	GTC	Silent
Val 232	C → G	GTC	GTG	Silent
Pro 236	C → G	CCC	GCC	-
Asn 238	T → G	AAT	AAG	Yes
Glu 246	G → C	GAG	CAG	-
Val 252	T → G	GTT	GTG	-
Gly 276	G → C	GGA	CGA	Yes

Out of 26 missing sites, 16 sites are $A \rightarrow C$ and $T \rightarrow G$ transversions with a frequency of 4.7 % (coloured blue in table), 9 sites are $G \rightarrow C$ and $C \rightarrow G$ transversions with a frequency of 4.1 % (orange in table), and 1 site is a $T \rightarrow C$ transition with a frequency of 17.5 % (green in table). Except the single $T \rightarrow C$ transition all other missing substitutions seem to be rare transversions occurring with an inherently low frequency (**Table 2. 2**). For 8 out of 26 sites, including the rare $T \rightarrow C$ transition, the missing amino acid is encoded by an alternative codon for which the mutation frequency is not rare.

Experimental design of the Alr library experimental evolution

To estimate the fitness of mutant residues relative to the wild-type residue we designed an experiment where *alr_{lib}* was allowed to grow for

Results Chapter 3.2

multiple generations in the presence and absence of selection pressure on Alr function.

To avoid bottlenecking the library, we ensured that we have more than 10 times the initial diversity of the library ($N_0 = 3.2 \times 10^8$ cells) at each stage of the experimental evolution experiment. We started the experiment with a volume of cells equivalent to $15 \times N_0$ i.e. 4.93×10^9 cells. Each day of the experiment we ensured the passage of $10 \times N_0$ cells. We decided to compare the effects of mutations in Alr under two different conditions (C_1 and C_2). C_1 , when the cells are under a continuous selection pressure for Alr activity; C_2 , when there is no selection pressure for Alr activity. Alr is the sole enzyme responsible for providing the essential D-alanine in *S. thermophilus* when absent in the growth medium. Thus without D-alanine in the medium, the Alr is under continuous selection pressure for activity. For condition C_1 , the cells were grown in medium devoid of D-alanine (orange colour in **Figure 2. 6 A**). For condition C_2 , the cells were grown in medium where D-alanine was exogenously added (Coloured in red in **Figure 2. 6 A**). The experiment was carried out in two replicates. The first day of the experiment (d_1) the cells were grown with D-alanine (C_2), consequently on d_2 two flasks were parallelly inoculated each for condition C_1 and C_2 , and were treated as replicates henceforth (**Figure 2. 6 B**). Cells were collected each day, lysed and the genome was purified and extracted as discussed above for initial library.

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

The PCR amplicons corresponding to days 1, 2 and 8 (nodes 2, 3, 4, 5, 6, 7, 27, 28, 29, 30 in **Figure 2. 6 B**) were deep sequenced.

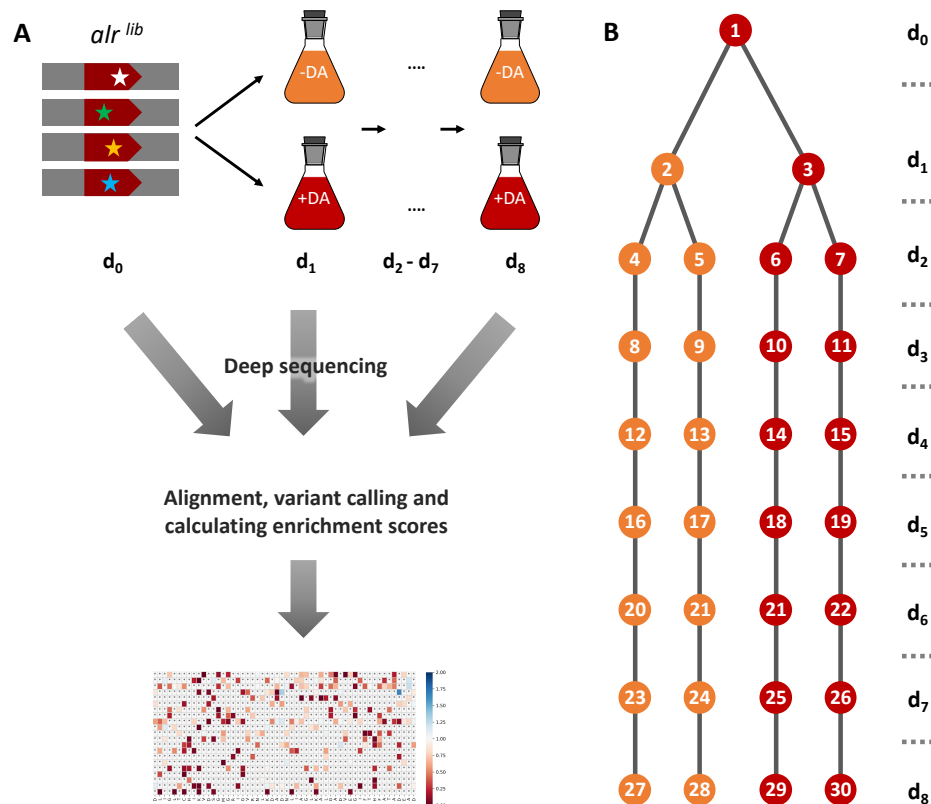


Figure 2. 6. Experimental plan for the prolonged experimental evolution of the Alr library. (A) The library was subjected to prolonged growth in media with and without D-alanine. Samples were taken at different time points and sequenced **(B)** Scheme explaining the replicates in this study.

Results Chapter 3.2

The next section is divided into two parts. First, all results observed with growth in condition C_1 will be presented and discussed. Subsequently drift in condition C_2 will be presented and discussed.

Part I : Alr library growth in condition C_1

Phenotypic characterization of evolving Alr library in condition C_1

To phenotypically characterize the experimental evolution, we retransformed the parent strain ($\Delta alr::P_{32-cat-P_{sp-oroP}}$) with the extracted and amplified *alr* gene library from the cells collected at each timepoint of the growth ($\Delta P_{32-cat-P_{sp-oroP}}::alr^{lib (day N)}$). We grew the transformed cells on plates of growth media with and without supplemented D-alanine. Mutants are considered to be active if they can complement the growth of the cell in growth media without supplemented D-alanine. Number of colony forming units (cfu) observed in each condition for three technical replicates were counted and averaged, subsequently to calculate the percentage of active mutants equation (1) was used. On average 1100 cfu were counted to minimize statistical variations due to sampling in cfu counting.

$$Alr \text{ percent active} = \left(\frac{cfu \text{ without } D\text{-alanine}}{cfu \text{ with } D\text{-alanine}} \right) \times 100 \quad (1)$$

As expected in condition C_1 we observe that active Alr mutants are rapidly selected in the library (**Figure 2. 7**) and by day 8 only active

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

mutants survive (0.2 % to 94.98 % ± 3.07). One might wonder how few inactive mutants survive the initial days of the experimental evolution in condition C_1 . Previous work by Satoshi Katsube et. al. has shown that in conditions of excess L-alanine, *E. coli* cells can secrete D-alanine and this secreted D-alanine can support the growth of a D-alanine auxotroph (Katsube et al. 2016). In *S. thermophilus*, there might be a similar secretion of D-alanine from Alr active cells growing in liquid culture, supporting the growth of cells with inactive or cells with low Alr activity.

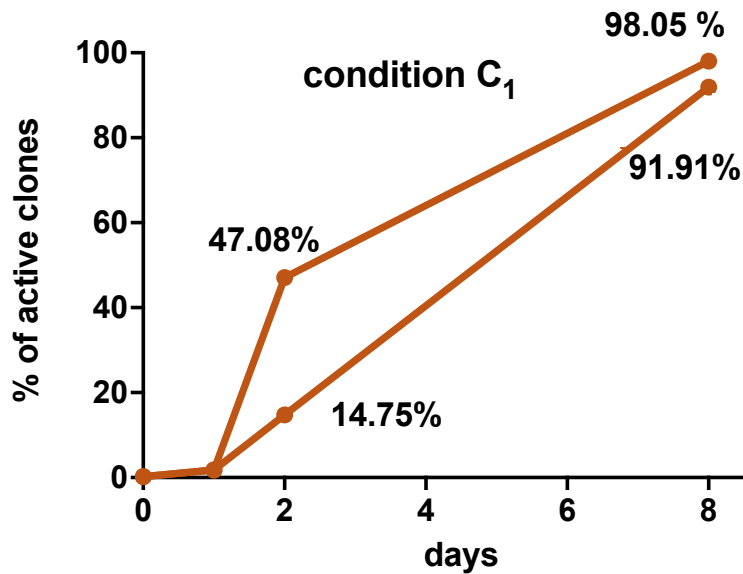


Figure 2. 7. Phenotypic characterization of the Alr library in condition C_1 . Back-transformation in condition C_1 demonstrates rapid selection for Alr activity yielding almost ~100% increase in percentage of active mutants. Number of colony forming units (cfu) observed in each

Results Chapter 3.2

condition for three technical replicates were counted and averaged, subsequently to calculate the percentage of active mutants equation (1) was used. On average 1100 cfu were counted to minimize statistical variations due to sampling in cfu counting.

Deep mutational scanning of experimental evolution in condition C_1

We utilize the results of deep sequencing to calculate enrichment ratios (E) of each wild-type to mutant amino acid substitution. For a site r , with wild-type residue x and mutant residue m , the total number of reads is given by N_r . Count of reads for m is given by $N_{m,r}$. A threshold of minimum 10 counts was set for $N_{m,r}$ values, to avoid high variance due to low counts.

Frequency of m at position r can be calculated by,

$$f_{m,r} = \frac{N_{m,r}}{N_r} \quad (2)$$

Similarly, for x at position r ,

$$f_{x,r} = \frac{N_{x,r}}{N_r} \quad (3)$$

If the frequency of m at timepoint 0 in the experimental evolution is $f_{m,r}^0$, frequency of m at timepoint t in the experimental evolution is $f_{m,r}^t$ while

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

frequency of x at timepoint 0 in the experimental evolution is $f_{x,r}^0$,

frequency of x at timepoint t in the experimental evolution is $f_{x,r}^t$.

The enrichment ratio for m at site r (E_m^r) can be calculated as,

$$E_m^r = \frac{\left(\frac{f_{m,r}^t}{f_{x,r}^t} \right)}{\left(\frac{f_{m,r}^0}{f_{x,r}^0} \right)} \quad (4)$$

If the value of E_m^r lies between 0 and 1 ($0 \leq E_m^r < 1$), m is purged; whereas if the value of E_m^r is greater than 1 ($E_m^r > 1$), m is enriched. If the value of E_m^r is equal to 1 ($E_m^r = 1$), m is neutral.

Enrichment ratios were calculated for each of the 1712 substitutions for experimental evolution in condition C_1 at timepoints d_0 , d_1 , d_2 and d_8 using equations (2), (3) and (4). Correlation between the replicates were estimated by computing the Pearson correlation coefficient for each pair of E_m^r values for the two replicates. We obtain a R value of 0.93 between the replicates for the condition C_1 (Figure 2. 8). For the remainder of this study, we have combined enrichment values by averaging the values across the two replicates.

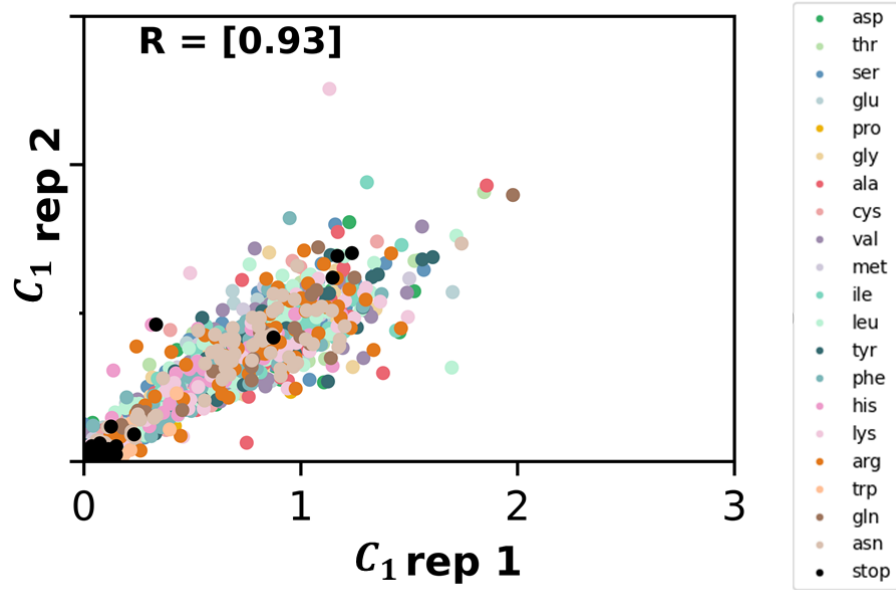


Figure 2. 8 Correlation of replicates for experimental evolution in condition C_1 . Correlation among the calculated enrichment ratios in the experimental evolution among the replicates. Each dot represents a pair of enrichment ratio values, the colours represent different codons after mutation as shown in the legend, and the R value represents the Pearson correlation coefficient. (Python script – Appendix C)

Average number of mutations before and after experimental evolution in condition C_1

We estimated the change in the average number of missense and nonsense mutations per gene through the experimental evolution. For condition C_1 , it decreases from 2.9 at d_0 to 1.4 at d_8 (**Figure 2. 9**),

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

indicating a general purge of mutations. The values shown are obtained by equation (5), where x^t is the average number of mutations/gene, N_m^t is the number of missense plus nonsense mutation reads at timepoint t , and N^t is the total number of reads at timepoint t .

$$x^t = \frac{N_m^t}{N^t} \times \text{length of gene} \quad (5)$$

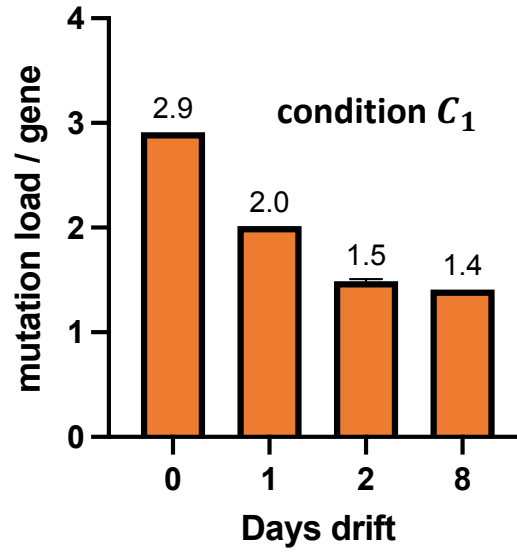


Figure 2. 9 The average number of missense + non sense mutations per gene across the timepoints of the experimental evolution in condition C_1 . The histogram represents the mean of replicates and error bars signify the SD of replicates.

Overview of selection/purge of mutations in condition C_1

Results Chapter 3.2

The combined enrichment values for $C_1 (E_m^r)$ were used to plot the selection/purge of amino acids at all sites. The enrichment values ($\ddot{E}_m^r$) were transformed for better visual representation in the sequence logo plot using equation (6) and (7).

$$\text{If } E_m^r \geq 1 \quad \ddot{E}_m^r = E_m^r - 1 \quad (6)$$

$$\text{If } E_m^r < 1, \quad \ddot{E}_m^r = 1 - E_m^r \quad (7)$$

This transformation of E_m^r values was to better visualise the extremes of selection and purge. These values were only used to generate the logo plot (**Figure 2. 10**), for all other analysis/figures untransformed E_m^r values are reported.

RaptorX server (Wang et al. 2016) was used to calculate the values for the secondary structure and relative solvent accessibility for the Alr residues. Relative solvent accessibility (RSA) for each residue is reported as either buried ($RSA < 25\%$), medium or exposed ($RSA > 75\%$). Secondary structure is reported as either helix, beta-sheet or loop. Finally, the *dmslogo* python package was used to visually represent the $\ddot{E}_m^r$ values as a sequence logo plot and to represent the secondary structure and relative solvent accessibility values as an overlay. The sequence logo plot displaying $\ddot{E}_m^r$ values as heights of 1712 substitutions for condition C_1 is shown in **Figure 2. 10**. We can see that for most sites, substitutions are strongly purged (large letters in red), a few sites have selected substitutions (large blue letters). These distribution of fitness effects of the substitutions will be systematically studied below.

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

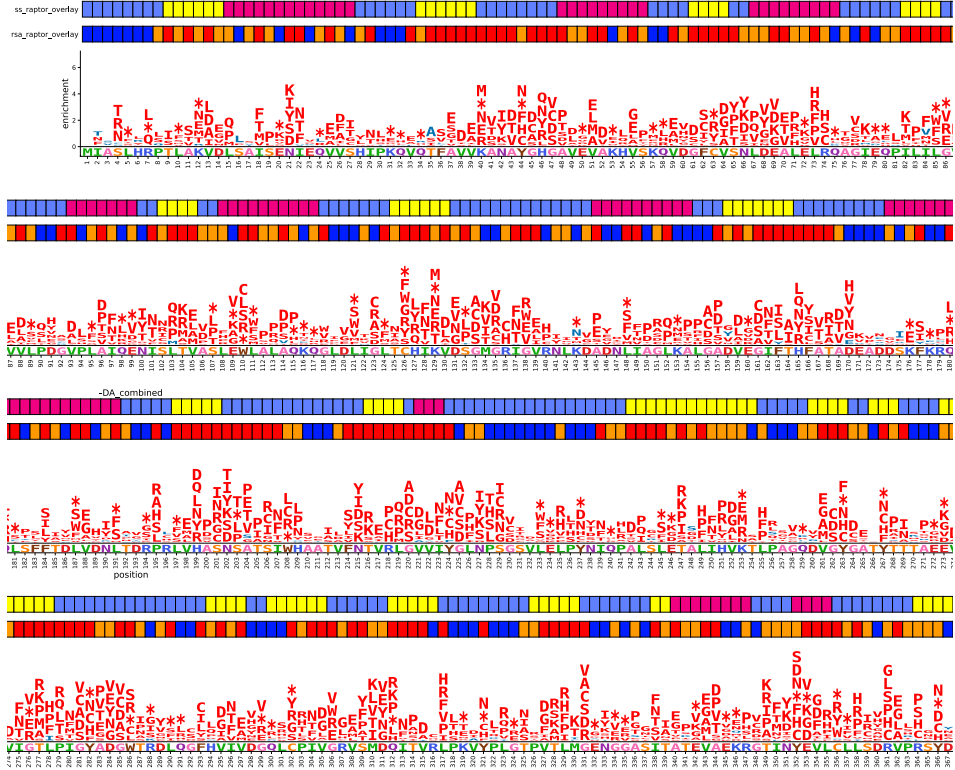


Figure 2. 10 Effects of substitutions in the sequence of AlrST subject to experimental evolution in condition C_1 as measured in the deep mutational scanning. E_m^r values were transformed to observe the larger differences more clearly. The transformed enrichment ($\ddot{E}_m^r$) values for each mutation is plotted as the height of the single letter code encoding the amino acid for condition C_1 . The height of each letter corresponds to the extent of selection or purge. The purged residues are coloured in red while the enriched residues are coloured in blue. The wild-type amino acid is displayed below the black line coloured by the amino acid functional groups. RaptorX was used to predict the relative solvent accessibility (rsa)

Results Chapter 3.2

and secondary structure (ss) of the residues, and these values are overlaid on the logo plot as coloured blocks (ss legend: alpha helices are coloured magenta, beta sheets are coloured yellow, turns are coloured pale blue ; rsa legend: blue signifies exposed residues, orange for middle residues and red for buried residues).

Distribution of fitness effects

The substitutions in condition C_1 will be systematically discussed now. Substitutions were separated into bins of 0.05 on the basis of the E_m^r values. 5 equally distributed class labels were overlaid on the graph,

Deleterious (del)	E_m^r value less than 0.4
Slightly deleterious (s.del)	E_m^r value between 0.4 and 0.8
Neutral (neu)	E_m^r value between 0.8 and 1.2
Slightly beneficial (s.ben)	E_m^r value between 1.2 and 1.6
Beneficial (ben)	E_m^r value greater than 1.6

Purged substitutions in condition C_1

Non sense mutations

In the initial library, 89 residues mutated to stop codons. 100 % of these residues have E_m^r values between 0 and 0.4, signifying that all stop codons were strongly purged in the experimental evolution for condition C_1 (Figure 2. 11).

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

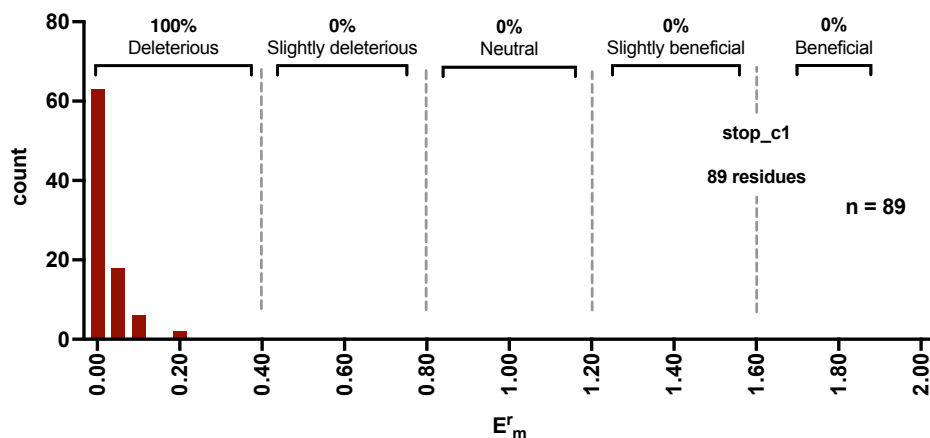


Figure 2. 11. Distribution of fitness effects of substitutions to stop codons in Alr experimental evolution in condition C_1 .

Active site

The 20 active site residues are listed in **Table 4**. 79% of substitutions in active site residues were deleterious (**Figure 2. 12**). 85% of deleterious substitutions are non-conservative.

A closer look at the essential active site residues K40, the lysine covalently bound to PLP, and Y263, the tyrosine acting as catalytic base, reveals some interesting observations. K40R substitution shows a much higher E_m^r value as compared to the other mutations (0.22 for K40R as compared to 0 for all others). Previous study with another PLP containing enzyme glutamate-1-semialdehyde aminotransferase from *Synechococcus* report that, while most substitutions abolish activity, Lysine to Arginine mutants are not inactive, but presents highly reduced activity (Grimm, Smith, and von Wettstein 1992). We hypothesize that, similarly, AlrSt K40R would have reduced activity. Regarding Y263, a

Results Chapter 3.2

mutation is also not completely purged (0.12 for Y263C), while others are purged (0 for Y263F or Y263Stop). Previous studies with alanine racemase mutants from *Geobacillus stearothermophilus* have shown a highly reduced residual activity for Tyrosine to Serine mutant, while Tyrosine to Phenylalanine show almost no activity (Watanabe, Yoshimura, et al. 1999). This suggests that AlrSt Y263C is also endowed with some residual activity.

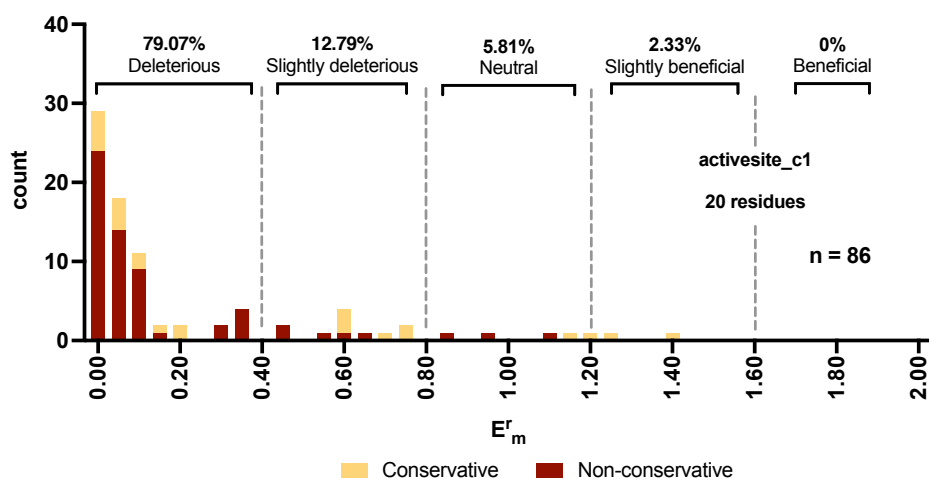


Figure 2. 12 Distribution of fitness effects of active site substitutions for Alr experimental evolution in condition C_1 . Conservative substitutions are displayed in yellow and non-conservative substitutions are displayed in red.

Relative Solvent accessibility (RSA)

The residues not involved in forming the active site or dimer interface were separated into three groups based on their RSA values. Overall the

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

tolerance and purge of substitutions are well correlated with RSA values (**Figure 2. 14**). As expected, 58% of substitutions in buried residues were deleterious. In the middle residues, the distribution of deleterious effects are more dispersed. 34% substitutions were deleterious, while 42% were slightly deleterious. For the exposed residues, only 13% of substitutions were deleterious while 85% were predominantly slightly deleterious or neutral. Since typically mutations in exposed residues are tolerated, the most purged ($E_m^r < 0.1$) substitutions were analysed to evaluate the cause of the purge. Evolutionary conservation of AlrSt positions was calculated from an alignment of 300 homologs identified by a JackHMMER search (Johnson, Eddy, and Portugaly 2010) on the Consurf-HSSP server (Glaser et al. 2005) using AlrSt as input. Using the evolutionary conservation and E_m^r values, we offer some hypothesis to explain the purged exposed residues,

- 4 substitutions (Q115P, A150P, A154P, T268P) to proline for residues involved in forming α -helices.
 - We hypothesise, since proline can disrupt α -helices, this could be destabilizing the protein.
- 7 substitutions to a hydrophobic residue.
 - Substitutions N100I, N215Y, N215I were not observed in any homologs. Substitutions I97F, E99V, N100Y, W208L were observed in homologs.

Results Chapter 3.2

- We hypothesise, since surface exposed hydrophobic residues could destabilize proteins, these substitutions could have been purged for this.
- G123C
 - In sequence alignments of homologs, all amino acids except Cysteine and Phenylalanine (G123F substitution not covered in initial library) are present at this position.
 - Maybe cysteine exposed on the surface could cause unintended reactivity by providing a free thiol group, or by binding to metal ions.
- W208
 - All substitutions in the library (W208R, W208C, W208L) were purged ($E_m^r < 0.03$).
 - Only the nitrogen side of the indole group is exposed to solvent while the aromatic ring is occupying a hydrophobic cavity. This explains the conservation.
- Arg74
 - All substitutions in the library (R74C, R74H, R74S) were purged ($E_m^r < 0.08$).
 - This residue is highly conserved in homologs.
 - The guanidinium group of the arginine is stabilised by four main chain carbonyls (**Figure 2. 13**), any other residue would likely

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

disrupt this interaction. Arg74 may be contributing significantly to protein stability.

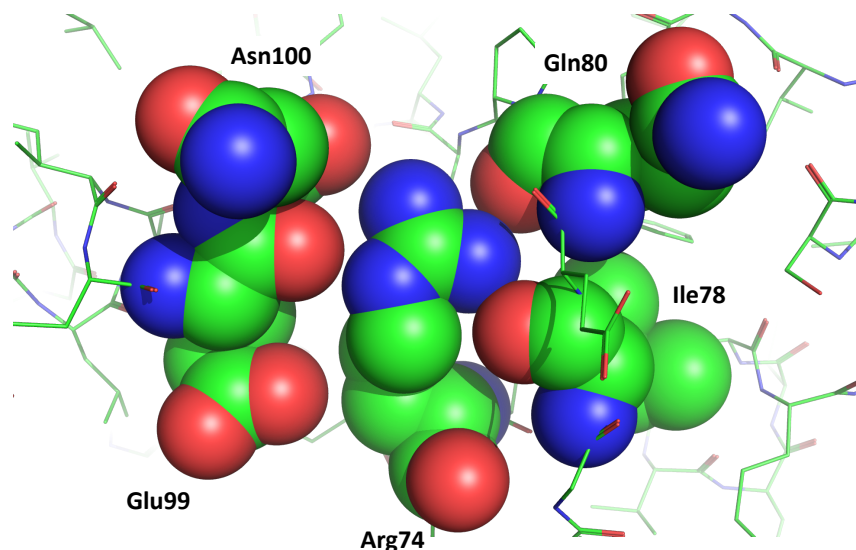


Figure 2. 13. Interactions between the guanidinium group of Arg74 and carbonyl carbons of residues Ile78, Gln80, Glu99 and Asn100. The residues are represented as spheres and the rest of the protein is represented as wireframe. Spheres are coloured by their element (Carbon, green; Nitrogen, blue; oxygen, red). Structure – AlrSt homology model.

- Tyr366
 - This position is fairly conserved in homologs. Tyr in 92% and Phe in 7.2% of homologs. The main chain is exposed but the side chain is located in a small hydrophobic cavity, explaining the conservation.

Results Chapter 3.2

- Conservative substitution are tolerated Y366F ($E_m^r = 0.99$), While other substitutions were purged (Y366H, Y366C, Y366N, Y366D).

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

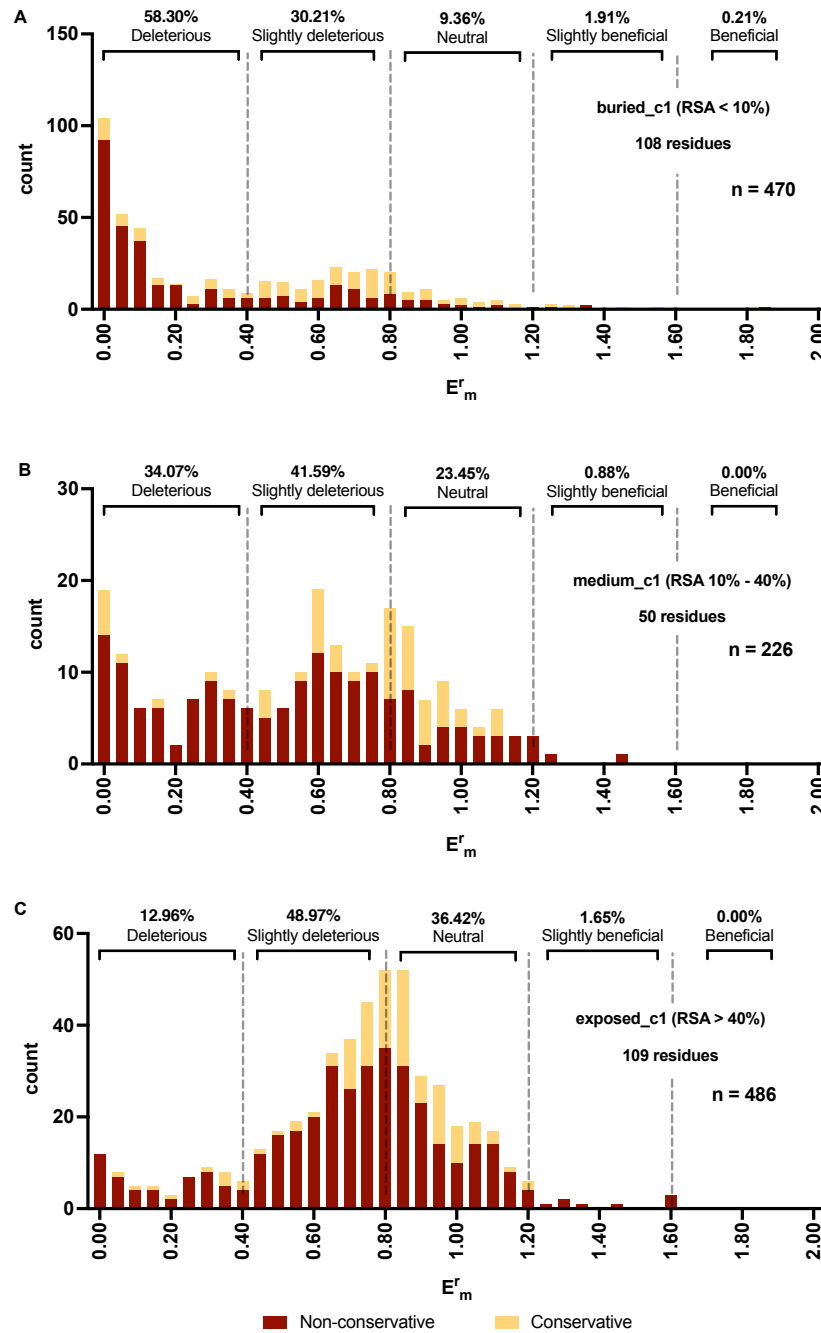


Figure 2. 14 Distribution of fitness effects of substitutions. (A) buried, RSA < 10%; (B) middle, RSA = 10% - 40%; (C) exposed, RSA > 40% residues for Alr experimental evolution in condition C_1 . Conservative substitutions are displayed in yellow and non-conservative substitutions are displayed in red.

RSA of interface residues

The residues at the dimer interface are expected to be conserved since the active sites are formed by residues from each monomer. To check if interface residues show a distribution of fitness effects that is different from other residues, interface residues were similarly separated into three groups based on their RSA values. 66% of buried interface residues were deleterious, 53% of middle interface residues were deleterious and 13% of exposed interface residues were deleterious (**Figure 2. 15**).

Overall, substitutions in buried and middle interface residues reveal a higher percentage of deleterious effects when compared to non-interface residues (**Table 2. 4**). The percentage of deleterious substitutions in buried residues is 8% higher in interface residues than non-interface residues, while the percentage of deleterious substitutions in middle residues is 19% higher in interface residues than non-interface residues. Substitutions in exposed residues have similar effects in the interface and non-interface residues ($\pm 0.30\%$).

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

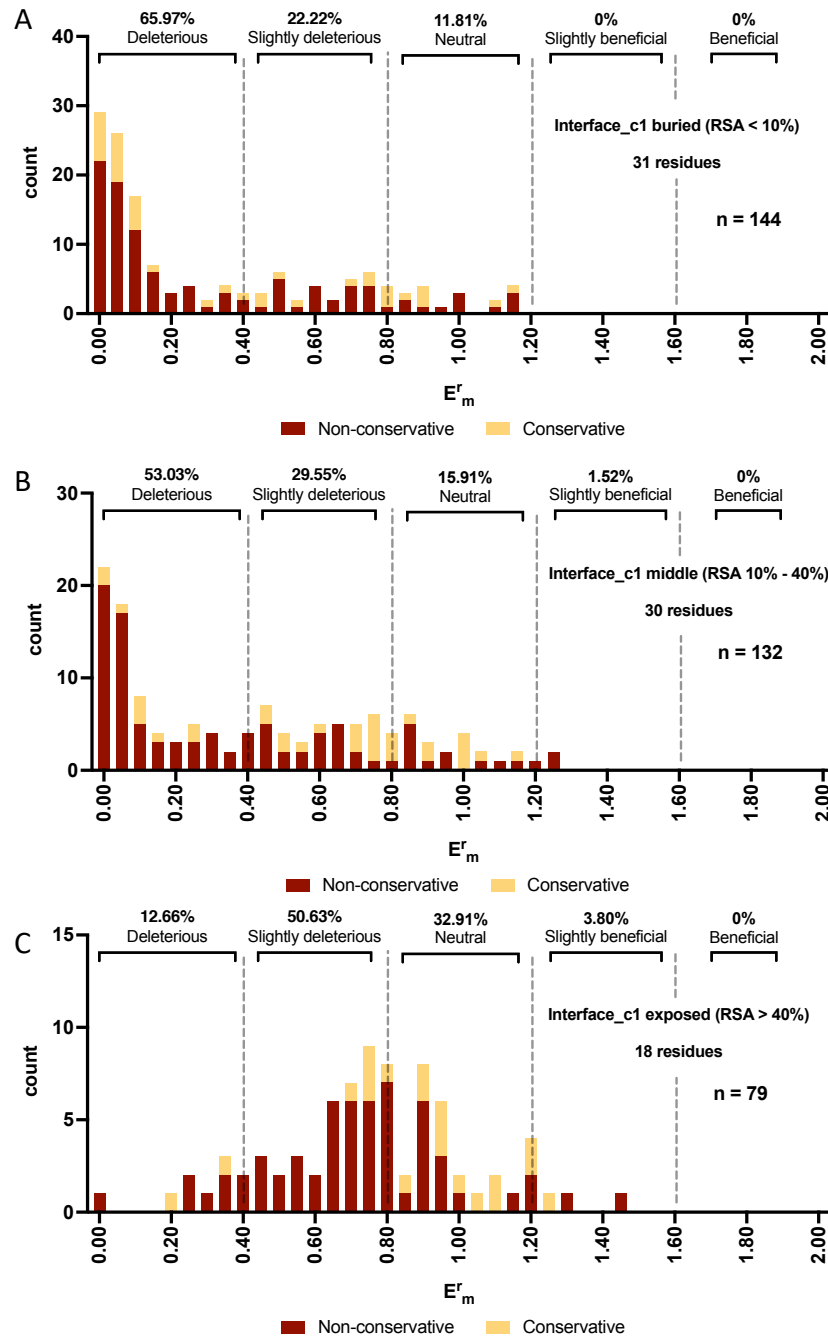


Figure 2. 15 Distribution of fitness effects of substitutions in residues involved in the dimer interface. (A) buried, RSA < 10%; (B) middle, RSA = 10% - 40%; (C) exposed, RSA > 40% residues for Alr experimental evolution in condition C_1 . Conservative substitutions are displayed in yellow and non-conservative substitutions are displayed in red.

Table 2. 4. Comparison of distribution of fitness effects of non-interface and interface residues.

RSA	residue type	n	del	s.del	neu	s.ben	ben
buried	non-interface	470	58.30%	30.21%	9.36%	1.91%	0.21%
	interface	144	65.97%	22.22%	11.81%	0%	0%
middle	non-interface	266	34.07%	41.59%	23.45%	0.88%	0.00%
	interface	132	53.03%	29.55%	15.91%	1.52%	0%
exposed	non-interface	486	12.96%	48.97%	36.42%	1.65%	0.00%
	interface	79	12.66%	50.63%	32.91%	3.80%	0%

General trends in distribution of fitness effects for condition C_1

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

Table 2. 5. Distribution of fitness effects for 1712 substitutions in Alr experimental evolution condition C_1 .

Type of residue	Number of mutations (n)		del	s.del	neu	s.ben	ben
stop codons	89	n	89	0	0	0	0
		%	100%	0%	0%	0%	0%
active site (20 residues)	86	n	68	11	5	2	0
		%	79%	12.79%	5.81%	2.33%	0%
interface (79 residues)	355	n	175	111	64	5	0
		%	49.30%	31.27%	18.03%	1.41%	0%
buried (108 residues)	470	n	274	142	44	9	1
		%	58.30%	30.21%	9.36%	1.91%	0.21%
medium (50 residues)	266	n	91	111	62	2	0
		%	34.07%	41.59%	23.45%	0.88%	0.00%
exposed (109 residues)	486	n	63	238	177	8	0
		%	12.96%	48.97%	36.42%	1.65%	0.00%
Total (366 residues)	1712	n	746	596	343	26	1
		%	43.57%	34.81%	20.03%	1.52%	0.06%

Selected substitutions in condition C_1

Results Chapter 3.2

The selected substitutions in condition C_1 will be systematically discussed now. The selected substitutions were divided into different groups based on the residue and the type of substitution. These residues are listed in **Table 2. 6**

Table 2. 6 Substitutions selected for Alr experimental evolution in condition C_1 .

Substitution	E_m^r	Cons
P380Q	1.89	
T35A	1.86	4
S370T	1.83	
Q381K	1.82	
S16L	1.62	3
K143N	1.60	3
S175I	1.59	1
I84V	1.57	7
H379Y	1.49	
D158Y	1.46	1
L278I	1.46	5
S375T	1.45	
I2T	1.44	7
A248S	1.43	5
G368D	1.42	
G374C	1.42	

Substitution	Erm	Cons
F163Y	1.30	6
P256S	1.30	1
S369F	1.29	
R196L	1.29	2
D367Y	1.28	1
P380L	1.28	
W377*	1.28	
F177Y	1.26	7
K346Q	1.25	1
V320A	1.25	1
Y224F	1.24	9
P8Q	1.24	6
Q259K	1.23	4
S370L	1.23	
R288K	1.23	9

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

D170E	1.41	9
S372R	1.41	
A204S	1.37	9
G376S	1.36	
P380A	1.36	
L103M	1.36	4
Q24L	1.34	2
G376D	1.34	
S375*	1.32	
A114V	1.31	5

Residues are coloured according to the type of substitution- residues of active site and entryway coloured in blue; interface residues coloured in green; residues of the strep tag coloured in black; substitutions resulting in exposed hydrophobic residues coloured in orange; conservative substitutions coloured in yellow; rest uncoloured. The residues are arranged by their E_m^r values. **Cons-** Consurf evolutionary conservation scores.

Residues selected in the C-terminal strep tag

There are 43 missense substitutions and 5 stop codon substitutions in the 17 residue strep-tag. The distribution of mutations is 46% neutral (20/43), 25% slightly beneficial (11/43), 20% slightly deleterious (9/43) and 7% beneficial (3/43). Substitutions to stop codons

Results Chapter 3.2

are 60% neutral (3/5) or 20% (2/5) selected. We hypothesise that there is a slight fitness cost to expressing the tag, even at physiological levels and the beneficial mutations might be compensatory (Figure 2. 16). Interestingly, stop codons in positions 375 ($E_m^r = 1.32$), 377 ($E_m^r = 1.27$) and 381 ($E_m^r = 1.19$) are selected while those in position 370 ($E_m^r = 0.85$) and 373 ($E_m^r = 0.62$) are purged, indicative that shortening the tag is not a general solution for improving the fitness.



Figure 2. 16. Substitutions selected in the C-terminal strep tag for condition C₁. Transformed enrichment ($\ddot{E}_m^r$) values for each mutation is plotted as the height of the single letter code encoding the amino acid. Selected residues are displayed in blue and purged residues are displayed in red. The wildtype residues are displayed below the black line.

Hydrophobic substitutions selected at exposed residues

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

At six exposed positions in Alr, mutations to hydrophobic residues were selected. This is interesting, since solvent exposed hydrophobic residues are known to destabilize proteins (**Table 2. 6**). One hypothesis is that these substitutions could have been selected to compensate for the effect of the tag.

Residues selected in the active site are clustered around the active site entryway

The entry to the active site of AlrSt consists of three layers of conserved residues starting from predominantly conserved hydrophobic residues near the LLP residue to non-conserved charged residues near the outside. The list of residues involved in forming the active site entry in AlrSt can be found in **Table 3**. D170E, R288K, A204S were selected after experimental evolution in condition C_1 (**Table 2. 6**) even though it is difficult to imagine that these mutations would improve alanine racemase activity.

D170E, R288K are present in the middle layer of the active site. They are both conservative substitutions since aspartic acid is substituted by glutamic acid having a longer side chain, on the other hand arginine is substituted by lysine having a less-bulky side chain. The mutations change the size of the active site entryway **Figure 2. 17**. One possible explanation could be that D170E, R288K (separately or together) improve or decrease a yet undefined promiscuous activity, and any reduction in native

Results Chapter 3.2

racemase activity in this trade-off is buffered by the nutrient excess in the rich M17 media.

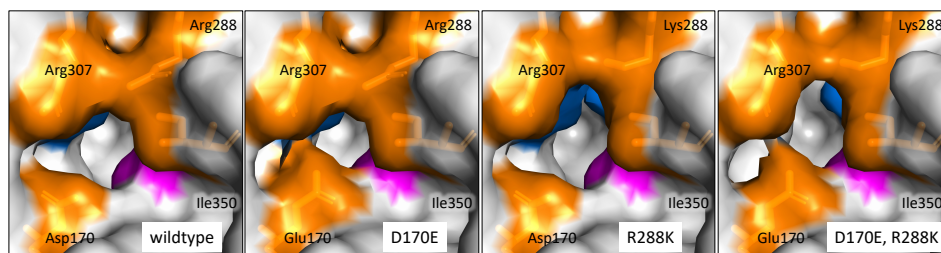


Figure 2. 17. Selected mutations D170E and R288K probably change size of active site entryway. AlrSt is represented as a Connolly surface to demonstrate the change in active site entry for the wild-type and the selected mutants. Catalytic residues are coloured in blue, inner entryway residues are coloured in pink, middle entryway residues are coloured in orange. Swiss model was used to model the wild-type and mutant enzyme structures as described in **Figure 2. 4**.

Correlation between enrichment values and evolutionary conservation of Alr

Next, we decided to compare the experimental enrichment values (E_m^r) in condition C₁ to the evolutionary conservation at the site. Most algorithms consider the conservation of residues at individual sites by using information extracted from multiple sequence alignment. PROVEAN is a predictor which considers the sequences at the residues which are neighbouring to a site in order to calculate a score (Choi et al. 2012).

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

Similarly, SIFT is a tool that utilizes sequence homology to predict whether a substitution would be deleterious or tolerated (P. C. Ng and Henikoff 2001). We used the sequence of AlrSt and the list of 1623 substitutions (1712 - nonsense substitutions) as input to calculate scores. We then compared our calculated E_m^r values to the scores obtained from PROVEAN and SIFT. We see a modest positive correlation ($R = 0.56$) between the experimental values and the PROVEAN scores, similarly for SIFT scores we see a positive correlation ($R = 0.42$) between the experimental values and the SIFT scores. The selection/purge observed in the experimental evolution correlates, at least moderately, to the observed diversity in Alr homologs. This shows that the fate of mutations also depends on the background sequence context of the protein in which it is placed. SIFT, which does not consider phylogenetic distances between the orthologs, showed lower correlation to the experimental values than PROVEAN (Figure 2. 18). This is in agreement with previous studies of fitness mappings of modification methylase HaeIII in *E. coli* (Rockah-Shmuel, Tóth-Petróczy, and Tawfik 2015) .

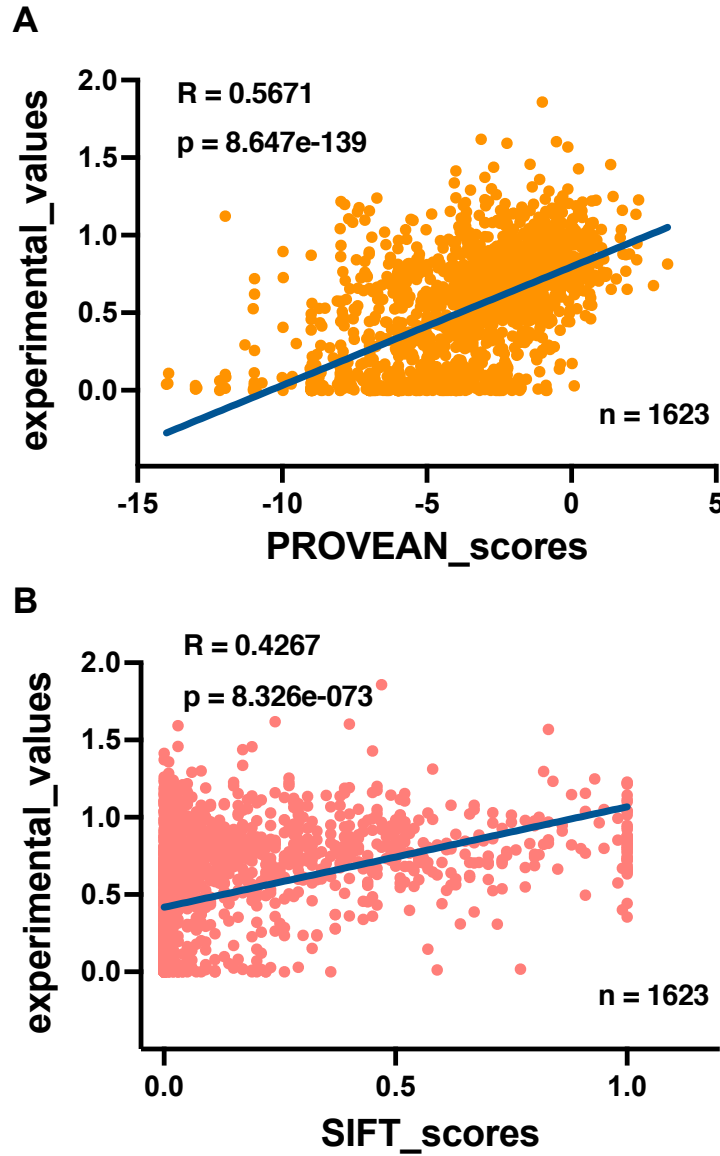


Figure 2. 18 Mutational effects for the experimental evolution in condition C_1 correlates with natural diversity of substitutions among Alr sequences. (A) Correlation of the experimental values with the PROVEAN

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

scores. (B) Correlation of the experimental values with SIFT scores. The R value is the Pearson correlation coefficient for 1623 pairs of values while the linear regression bars represent 95% confidence ratios. The *p*-values are calculated from a two-tailed test.

Mutations implicated in clinical isolates of D-cycloserine (DCS) resistance in Alr are under functional selection

Next, we chose to examine the mutations implicated in conferring DCS resistance in common clinical isolates and laboratory screening of *M. tuberculosis* strains. AlrSt and Alr^{MT} share 34% sequence identity with a well conserved active site. Most commonly implicated mutations are M320T (M310 in AlrSt), D322N (D311 in AlrSt), Y364D/C (Y352 in AlrSt), K157E (K129 in AlrSt) and R397G/L (R361 in AlrSt) (Chen et al. 2017; Desjardins et al. 2016; Evangelopoulos et al. 2019; Nakatani et al. 2017). Additionally, mutations L113R (V88 in AlrSt), T155A (H127 in AlrSt), S261N (N227 in AlrSt) have also been mentioned but since these amino acid positions are not conserved in AlrSt they were not included in the analysis. These Alr variants are also present in the common *M. tuberculosis* lineages. Mutations M320T, Y364D, R397G are present in strains belonging to the lineage 4 (Euro-American); M320T and Y364D/C are present in the lineage 2 (Beijing); while R397L and Y364D are present in lineage 1 (East African Indian). To facilitate further analysis, for the remainder of the study we will be referring to the amino acid positions from AlrSt (denoted in parenthesis earlier). Previous studies with purified

Results Chapter 3.2

Alr mutants from *M. tuberculosis* describe large reduction in turnover rates for the M310T and R361L mutants (Nakatani et al. 2017) and found the D311N mutant to be 20-fold less active than *wild-type* Alr^{MT} (Evangelopoulos et al. 2019). **Table 2. 7** lists the day 8 E_m^r values for the list of specific mutations discussed above.

Table 2. 7. Alr mutations implicated in drug resistance for clinical strains of *M. tuberculosis* (the positions correspond to positions in AlrSt)

Mutation	E_m^r in C_1
K129E	0.02
M310T	0.06
D311N	0.10
Y352D	0
Y352C	0.06
R361G	0
R361L	0

These mutations are strongly deleterious ($E_m^r < 0.2$) to alr function. Several mutually non-exclusive hypothesis could be imagined to explain this,

- a. Mutations involved in creating D-cycloserine resistance are tolerated by nutritional ‘buffering’ due to D-alanine shared within biofilms. It has been demonstrated in *Bacillus subtilis* that

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

mutualism within biofilms allows sharing of D-alanine. Clonally deleterious *alr* knockdowns, can survive in co-culture with D-alanine prototrophs (Arjes et al. 2021). We know that *M. tuberculosis* creates biofilms to help it escape the effects of antibiotics (Brennan 2017; Kulka, Hatfull, and Ojha 2012; Richards et al. 2019).

- b. Deleterious effects of these mutations are compensated by accompanying mutations that restore completely or partially *alr* activity.
- c. These mutations are deleterious at native expression levels, but total activity is increased by overexpression of the impaired enzyme. Indeed, many clinical strains of *M. tuberculosis* carry a deletion in the *alr* repressor (Evangelopoulos et al. 2019).

Overall, it seems that most of the D-cycloserine resistance mutations seem to be associated with a strong fitness-cost. Thus, prevalence of these mutations globally may be a consequence of widespread antibiotic use, additionally warning us to necessitate parsimonious use of antibiotics.

Discussion (Alr experimental evolution condition C₁)

Ryan et al. have shown that, in *S. cerevisiae*, cytosolic levels of proteins expressed under common yeast plasmids have considerable

Results Chapter 3.2

variation from cell to cell (Ryan et al. 2014). This brings into light the often overlooked assumption in experimental mappings that each variant is competing equally. Recently, multiple studies (Dandage et al. 2018; Jiang et al. 2013; Rockah-Shmuel, Tóth-Petróczy, and Tawfik 2015) have experimental systems where variants are expressed at physiological expression levels. But their system still relies on low copy number plasmids and, to our knowledge, variance in expression level has not been documented with such plasmids. Recently, studies have developed methods to integrate mutants directly at the chromosome, Ryan et al. have used CRISPR as a tool to integrate mutants in the chromosome of *E. coli* (1600 mutants) (Ryan et al. 2014), while Noda-Garcia et al. from Tawfik's team have integrated *gudB* mutants in the chromosome of *B. subtilis* (320 mutants) (Noda-García et al. 2019). Though, in both these methods the size of the library was small. Dorrazehi et. al. demonstrated the use of lambda red recombination in combination with CRISPR-cas9 to create libraries in *E. coli* of the range 10^5 to 10^6 mutants (Dorrazehi et al. 2020). However, to our knowledge, chromosomal libraries of the scale used in this study (10^8 mutants) have never been demonstrated before. In our experimental system, we introduce an elegant solution by cloning the variant library at the native chromosomal locus, owing to the high natural transformability of our host organism *Streptococcus thermophilus* we can easily obtain up to 10^9 transformants.

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

We employed this system to replace the native gene by a library of mutants by a 2-step *cat-oroP* scarless counter-selection strategy to create a large library of 3.29×10^8 variants. This system allows us to test the fitness effect of mutations while the enzyme is being expressed at exactly the native expression conditions. We demonstrate that Alr activity is essential to cell growth. Consequently, this library was simultaneously grown during 8 days in the absence of D-alanine in the medium.

The distribution of fitness effects for all the 1712 AlrSt substitutions after 8 days of growth in condition C_1 reveals interesting trends (**Table 2. 5**), 44% (746/1712) of all missense and nonsense substitutions were strictly deleterious with the threshold of $E_m^r < 0.4$. 35% (596/1712) mutations have a slightly deleterious effect with a threshold of E_m^r 0.4 - 0.8. Slightly deleterious substitutions could be mutations which may be tolerated in single mutants, but purged when they are present together in a single copy. As we see in the Poisson distribution in (**Figure 2. 5. C**). 16% of variants can be expected to be single mutants, while 79% of the variants in the library can be expected to be carrying > 1 mutations. We know that often destabilizing mutations will have additive effects (Gregoret and Sauer 1993; Wells 1990). Thus, a slightly deleterious substitution will be purged more rapidly when associated with other deleterious mutations. The evolution of our library with an average of 3 mutations per gene is therefore not fully comparable to what happens in nature but we can assume that the main difference will be in the rate of mutations evolution

Results Chapter 3.2

rather than in the identities of purged or selected substitutions. Repeating the experiment with a library designed with a lower mutational load could question this assumption but may require longer time and/or higher depth in sequencing. Additionally, we have overlooked epistatic effects of coupled mutations. It is not our intention to suggest that epistatic interactions are not important. However, it is beyond the scope of this study due to the short read Ion-torrent sequencing we used, to consider the epistatic effects of mutations.

About 20% (343/1712) of substitutions were neutral, 1.5% (26/1712) substitutions were slightly beneficial and 0.06% (1/1712) substitutions were beneficial. Interestingly at 6 exposed positions, hydrophobic substitutions were selected. Our hypothesis is that they may have been selected to compensate for the fitness cost associated with the tag. Although determining the nature of the fitness advantage granted by these substitutions needs further analysis, a plausible explanation would be a hydrophobic interaction that is moving the tag away from its site of interference. From an evolutionary point of view, such compensatory mutations may mimic natural mechanisms of protein evolution following sense mutation of a stop codon that increases the size of a protein with a random C-terminal extension.

We also see that the effect of the substitutions are strongly correlated to the RSA of the residues (**Figure 2. 14**). Interestingly, substitutions in buried and middle interface residues reveal a higher

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

percentage of deleterious effects than non-interface residues (**Table 2. 4**). We see that this is true for the core of the interface (buried, middle RSA) but not for the rim of the interface (exposed RSA). This result suggests that mutations affecting interfacial interactions may be in-general more deleterious than mutations affecting the hydrophobic packing of intra-subunit core structure.

The percentage of deleterious substitutions in buried residues is 7.6% higher in interface residues than non-interface residues, while the percentage of deleterious substitutions in middle residues is 19% higher in interface residues than non-interface residues. Substitutions in exposed residues have similar effects in the interface and non-interface residues ($\pm 0.30\%$).

Interestingly, substitutions D170E, R288K and A204S in the active site entryway were selected. A204 although not directly situated in the active site entryway is also located proximally. D170 and R288 are residues maintaining the active site entryway, and analysis of the structure (**Figure 2. 17**) suggests that these substitutions could create reduced access to active site. Since we do not expect these mutations to positively affect alanine racemase activity, our hypothesis is that selection pressure is acting on a promiscuous activity. Previous literature reported widespread promiscuity among alanine racemases. Alr from *Thermoanaerobacter tengcongensis* (Alr^{Tt}) was shown primarily to have racemase activity towards L-alanine, but relatively reduced activity

Results Chapter 3.2

towards L-arginine (7.9 %), L-methionine (5.9 %), L-lysine (5.1 %), L-serine (4.0 %) and L-valine (3.1 %) (Xue et al. 2013). In two other studies carried out with Alr from *Geobacillus stearothermophilus* and *Escherichia coli* (Patrick, Weisner, and Blackburn 2002; Soo et al. 2016) mutations in another inner active site residue Y354N in Alr^{GS} and Y274F in Alr^{EC} (Tyr352 in AlrSt) was found to be involved in active-site reshaping leading to promiscuous serine racemase (*G. stearothermophilus*) and cystathionine β -lyase activity (*E. coli*). The promiscuous Alr^{Tt} was also observed to have a histidine in the corresponding position 359 (Tyr352 in AlrSt), whereas, in our library, both mutations Y352N and Y352F were purged. Our experimental system has fixed parameters (1) native chromosomal expression and (2) high substrate concentration (large excess of L-alanine and other amino acids in rich M17 medium). Mutations acting on the substrate entryway are expected to have a negative impact on both Alr and any possible promiscuous activities, potentially decreasing enzyme K_m . Since M17, the growth medium of the experimental evolution, is concentrated in all amino acids, a decrease in K_m for L-alanine may not be associated with any fitness problem while any potential deleterious promiscuous activity, such as the production of toxic D-amino acids, could be reduced. Purification and characterization of mutants are currently ongoing.

Overall, as our experimental setup involves competition between mutants, we observe the effect of mutations directly on cellular growth

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

rate rather than enzyme activity. Since natural selection involves multiple selection pressures acting simultaneously including optimization of flux through the metabolic network, rather than selection for a single enzyme activity, this experimental design provides a more nuanced result of the effect of mutations on cellular fitness. It may also help us to better understand the advantages but also the potential problems that are associated with promiscuous activities. Indeed, an enzyme is usually seen as a very efficient catalyst for a given function but it should also be described as a secured catalyst that must not act on other metabolites. For the molecular evolution of some enzymes at least, this may create strong selection pressures.

Results Chapter 3.2

Results for drift in condition C_2 will be presented and discussed now

Part II : Alr library drift in condition C_2

Phenotypic characterization of drift of Alr libraries in condition C_2

Transformations were prepared as described for condition C_1 , and similarly equation (1) was used to calculate the percent of active clones. Surprisingly in condition C_2 , we see slower but still an increase in the percentage of active mutants (0.3 % to 2.8 % ± 0.4) (**Figure 2. 19**). This is interesting since, a priori, there is no selective pressure for active Alr variants. This will be discussed with the rest of the results at the end.

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

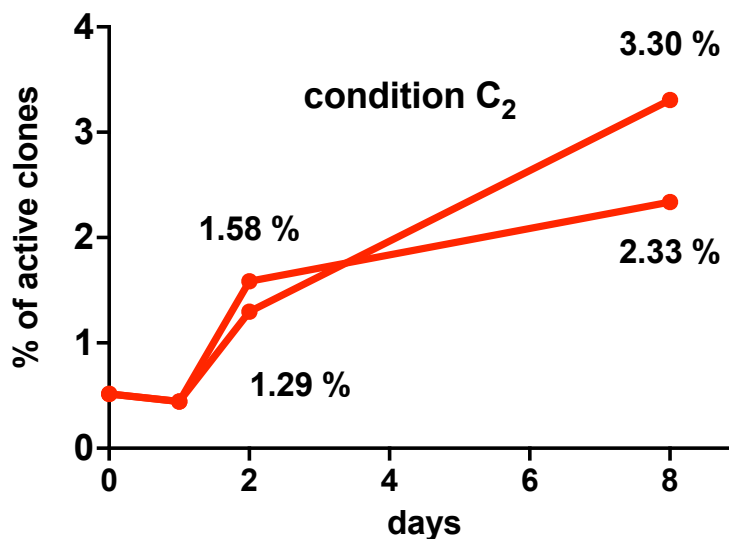


Figure 2. 19. Phenotypic characterization of drifted Alr library in condition C_2 . Back-transformation in condition C_2 yields an almost ~10% increase in percentage of active mutants. Number of colony forming units (cfu) observed in each condition for three technical replicates were counted and averaged, subsequently to calculate the percentage of active mutants equation (1) was used. On average 1100 cfu were counted to minimize statistical variations due to sampling in cfu counting.

Deep mutational scanning of drifted Alr library in condition C_2

Enrichment ratios (E_m^r) were calculated (as described for condition C_1) for each of the 1712 substitutions for drift in condition C_2 at timepoints d_0 , d_1 , d_2 and d_8 using equations (2), (3) and (4).

Results Chapter 3.2

Correlation between the replicates were estimated by computing the Pearson correlation coefficient for each pair of E_m^r values for the two replicates. We obtain a R value of 0.64 between the replicates for the condition C_2 (Figure 2. 20). As expected without absolute selection for enzyme activity, the correlation is weaker compared to replicates in C1 condition. Nevertheless, while the absence of correlation would have been an indicator of a fully neutral drift, the observed correlation suggests that selective pressures are acting and this is also in agreement with the unexpected selection of active clones. For the remainder of this section, we have combined enrichment values by averaging the values across the two replicates.

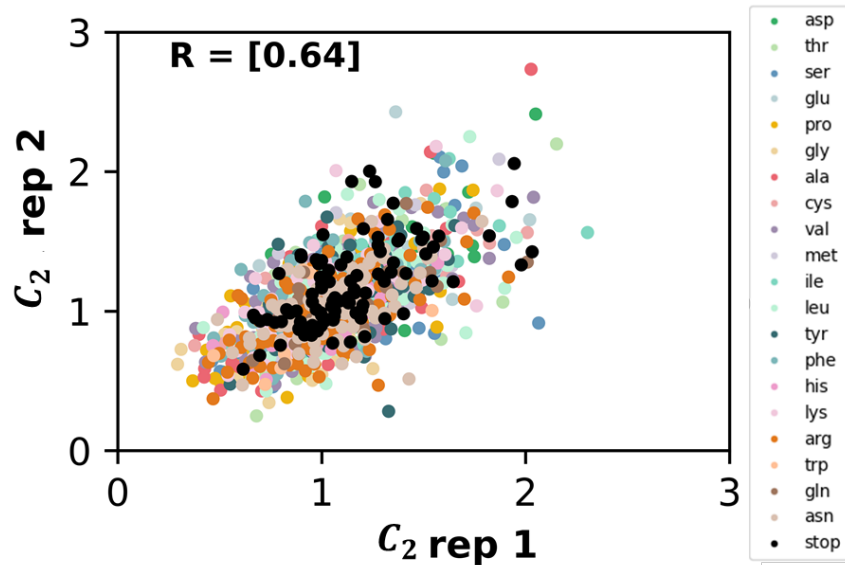


Figure 2. 20 Correlation of for drift in condition C_2 . Correlation among the calculated enrichment ratios in the drift among the replicates. Each

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

dot represents a pair of enrichment ratio values, the colours represent different codons after mutation as shown in the attached legend, and the R values represent Pearson correlation coefficients (Python script – Appendix C).

Average number of mutations before and after drift in condition C_2

As for condition C_1 , the average number of missense + non sense mutations/gene (x^t) for condition C_2 was calculated using equation (5). x^t for condition C_2 stays almost constant, increasing slightly from 2.9 at d_0 to 3.0 at d_8 (**Figure 2. 21**).

Since we know that the percentage of active clones is increasing, this result suggests that different selection pressures are acting on the library. Indeed, activity selection should be associated with a general purge of mutations. Therefore, a compensating pressure is probably acting by selecting mutants with a higher mutational load.

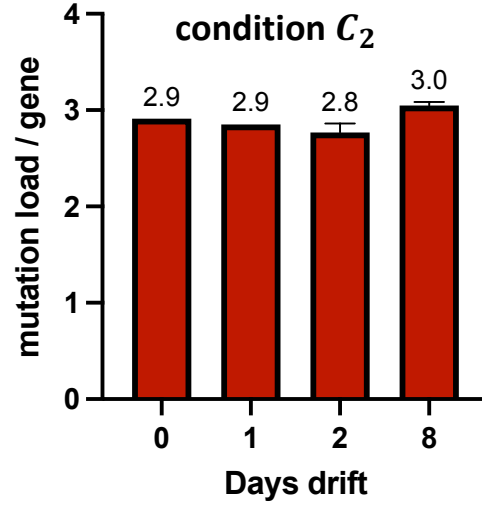


Figure 2. 21 The average number of missense + non sense mutations per gene across the timepoints of the drift in condition C_2 . The histogram represents the mean of replicates and error bars signify the SD of replicates.

Overview of selection/purge of mutations in condition C_2

As described for C_1 , the combined enrichment values for C_2 (E_m^r) were used to plot the selection/purge of amino acids at all sites. The enrichment values ($\ddot{E}_m^r$) were transformed for better visual representation in the sequence logo plot using equation (6) and (7).

The sequence logo plot displaying $\ddot{E}_m^r$ values as heights of 1760 substitutions (1712 CDS substitutions + 48 substitutions in strep tag) for condition C_2 is shown in **Figure 2. 22**. For most sites, selection (letters in

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

blue) or purge (letters in red) of substitutions do not show any clear predominance. The purged and selected substitutions will be systematically studied below.

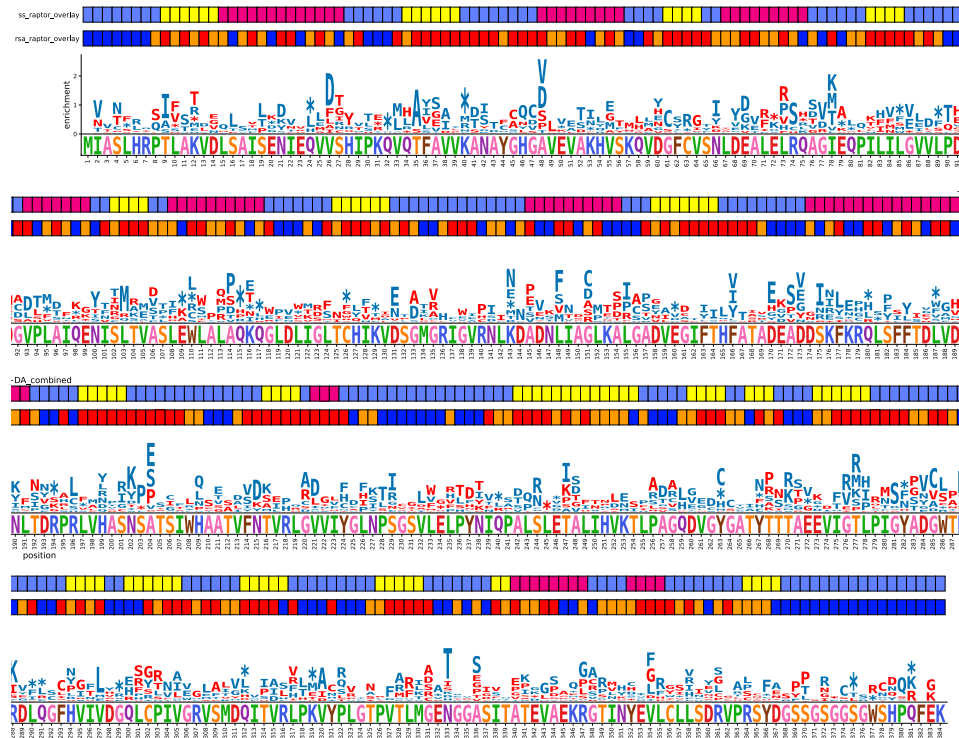


Figure 2.22 Effects of substitutions in the sequence of AlrSt subject to drift in condition C₂ as measured in the deep mutational scanning. The transformed enrichment ($\tilde{E}_m^r$) values for each mutation is plotted as the height of the single letter code encoding the amino acid for condition C₂. Purged residues are coloured in red, while selected residues are coloured in blue. The wild-type amino acid is displayed below the black line

Results Chapter 3.2

coloured by the amino acid functional groups. RaptorX was used to predict the relative solvent accessibility (rsa) and secondary structure (ss) of the residues, and these values are overlayed on the logo plot as coloured blocks (ss legend: alpha helices are coloured magenta, beta sheets are coloured yellow, turns are coloured pale blue ; rsa legend: blue signifies exposed residues, orange for middle residues and red for buried residues).

Distribution of fitness effects of substitutions in condition C_2

The substitutions in condition C_2 will be systematically discussed now. Substitutions were separated into bins of 0.05 on the basis of the E_m^r values. 5 equally distributed class labels were overlaid on the graph,

- | | | |
|------|------------------------------|-----------------------------------|
| i. | Deleterious (del) | E_m^r value less than 0.4 |
| ii. | Slightly deleterious (s.del) | E_m^r value between 0.4 and 0.8 |
| iii. | Neutral (neu) | E_m^r value between 0.8 and 1.2 |
| iv. | Slightly beneficial (s.ben) | E_m^r value between 1.2 and 1.6 |
| v. | Beneficial (ben) | E_m^r value greater than 1.6 |

Non sense mutations

Predominantly substitutions to stop codons were neutral (62%) or slightly beneficial (27%) and minority of substitutions were selected (4.4%) (**Figure 2. 23**). To check whether the benefit of stop codon substitution was related to its position in AlrSt we checked for any relation between position and E_m^r values, but no relation was observed.

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

Interestingly for the selected substitutions, the highest E_m^r value of 2 was observed for the catalytic Lys40 → Stop substitution.

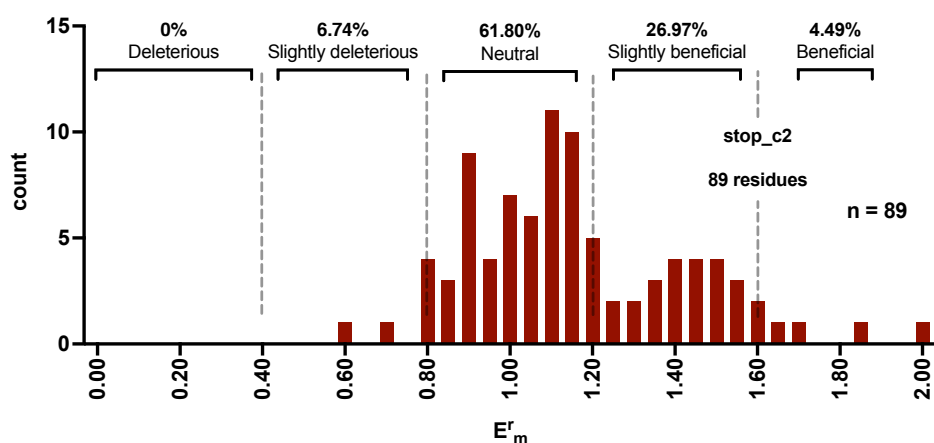


Figure 2. 23. Distribution of fitness effects of substitutions to stop codons in Alr drift in condition C_2 .

Active site

Predominantly substitutions were neutral (67%), slightly beneficial (15%) or slightly deleterious (10%) and minority of substitutions were beneficial (5.8%). Interestingly two out of five selected substitutions, R288K and D170K were also selected in condition C_1 . These mutations might be involved in improving or decreasing a promiscuous activity. Some hypotheses will be discussed later on.

- R288K ($E_m^r = 1.86$): Involved in middle layer of active site entryway, substitution could modulate enzyme active site entryway.

Results Chapter 3.2

- D170E ($E_m^r = 1.83$): Involved in middle layer of active site entryway, substitution could modulate enzyme active site entryway.
- S203P ($E_m^r = 1.72$): Situated near the inner active site entryway, substitution could change the available volume in the active site.
- Y263C ($E_m^r = 1.69$): Catalytic residue, substitution could allow bulkier substrates to be accommodated in the enzyme active site but racemase activity will be impaired.
- V221D ($E_m^r = 1.67$): Situated near the inner active site entryway, interacting with Y352.
- G220A ($E_m^r = 1.60$): Situated near the inner active site entryway, interacting indirectly with M310, Y352 though Y44.

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

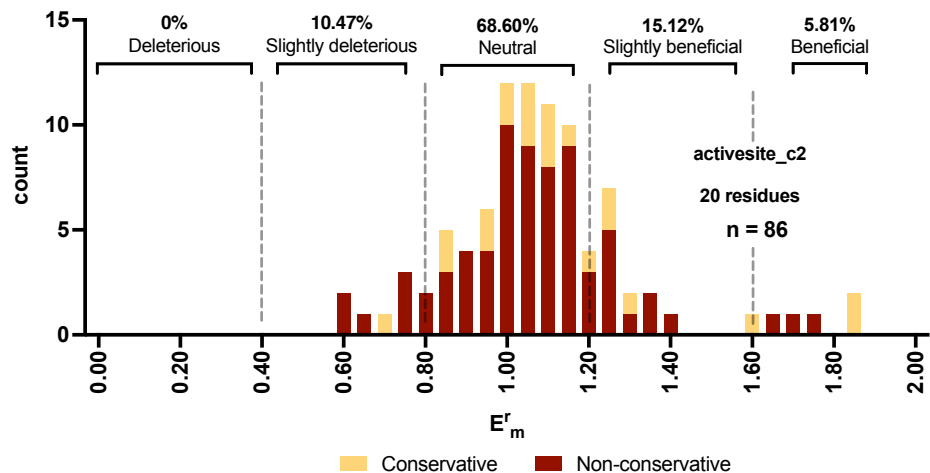


Figure 2. 24 Distribution of fitness effects of active site substitutions for Alr drift in condition C_2 . Conservative substitutions are displayed in yellow and non-conservative substitutions are displayed in red.

Interface residues

68% substitutions were neutral. 12% mutations were slightly deleterious, 17% mutations were slightly beneficial and 2% mutations were beneficial (**Figure 2. 25**).

Results Chapter 3.2

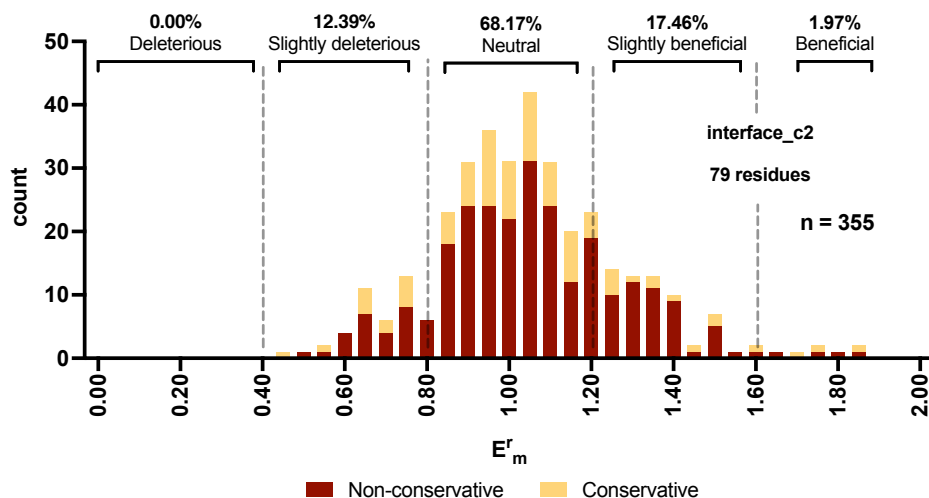


Figure 2. 25 Distribution of fitness effects of substitutions in residues involved in the dimer interface for Alr drift in condition C_2 .

Buried, middle and exposed residues

The residues not involved in forming the active site or dimer interface were separated into three groups based on their RSA values. In contrast with drift in condition C_1 , tolerance and purge of substitutions are not correlated with RSA values (**Figure 2. 26**). Most substitutions have neutral E_m^r values centred around 1.

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

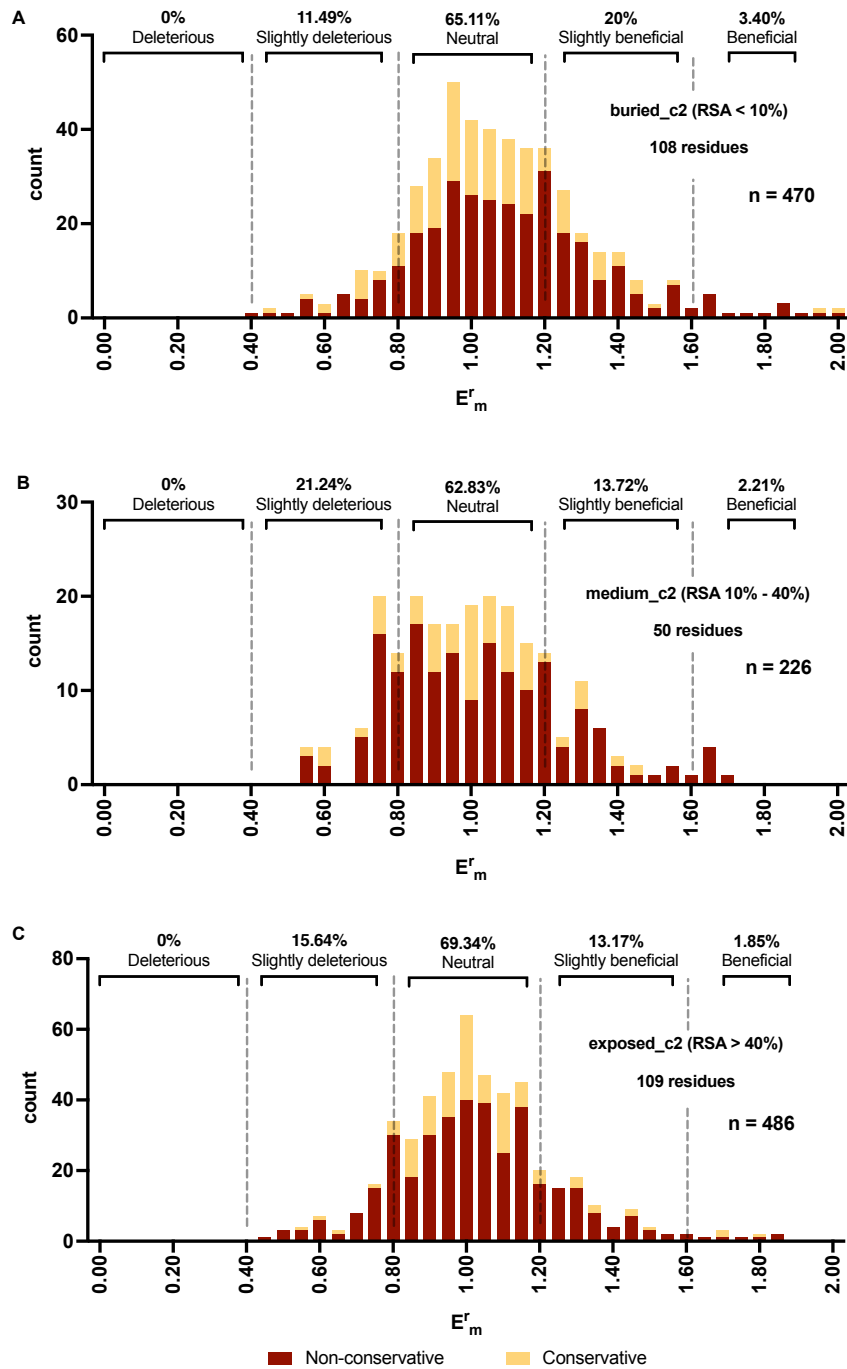


Figure 2. 26 Distribution of fitness effects of substitutions in (A) buried, RSA < 10%; (B) middle, RSA = 10% - 40%; (C) exposed, RSA > 40% residues for Alr drift in condition C_2 . Conservative substitutions are displayed in yellow and non-conservative substitutions are displayed in red.

General trends in distribution of fitness effects for condition C_2

Table 2. 8. Distribution of fitness effects for 1712 substitutions in Alr drift condition C_2 .

Type of residue	Number of mutations (n)		del	s.del	neu	s.ben	ben
stop codons	89	n	0	6	55	24	4
		%	0%	6.74%	61.80%	26.97%	4.49%
active site (20 residues)	86	n	0	9	59	13	5
		%	0%	10.47%	68.60%	15.12%	5.81%
interface (79 residues)	355	n	0	44	242	62	7
		%	0%	12.39%	68.17%	17.46%	1.97%
buried (108 residues)	470	n	0	54	306	94	16
		%	0%	11.49%	65.11%	20.00%	3.40%
medium (50 residues)	266	n	0	48	142	31	5
		%	0%	21.24%	62.83%	13.72%	2.21%
exposed	486	n	0	76	337	64	9

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

(109 residues)		%	0%	15.64%	69.34%	13.17%	1.85%
Total (366 residues)	1712	n	0	237	1141	288	46
		%	0%	13.84%	66.65%	16.82%	2.69%

Discussion (Alr drift condition C₂)

As expected, drift in condition C₂ presents a completely different picture as opposed to condition C₁. Although some common developments appear in both selections.

Phenotypic characterization of the drifting population shows a moderate increase (~10%) in the percentage of active clones in the library, this was surprising since we do not select for activity. At this stage, several non-mutually exclusive hypotheses may be proposed to explain this: (i) active clones still provide a slight fitness advantage associated with alanine racemization, possibly because alr contribute to alanine homeostasis in the cell as shown for *Bacillus subtilis* (Sidiq et al. 2020); (ii) active clones are selected because at least some of them are associated with a beneficial promiscuous activity; (iii) inactive clones or at least a fraction of them are purged because they are associated with a higher fitness cost compared to active clones; an attractive hypothesis is that alr unfolding may be associated with a stress in the bacteria, even at physiological level of expression.

Results Chapter 3.2

This increase is not really detected in the NGS data since active site mutations are not significantly purged at the library level. This indicates that this selection for activity is hidden by other selection pressures acting on the library. Since the increase in percentage of active clones is more or less linear between day 0 and day 8 (**Figure 2. 19**), the drift should be continued for a longer time to determine if the evolution will lead to a fully active library or if the trend is transient or reaching an intermediate plateau.

Overall, we observe that mutations are purged or selected to more or less the same extent, in agreement with the observation that the average number of mutations stays mostly constant (2.9 at d_0 to 3.0 at d_8) (**Figure 2. 21**) suggesting that the genes are mostly drifting neutrally. Although, drift in condition C_2 is moderately correlated between the replicates (**Figure 2. 20**), indicative of a selection pattern, the pattern of selection seems unpredictable at the amino acid level. This suggests that protein independent pressures could be acting at the DNA or RNA level, and may be dominating the selection. Comparison of the rate of nonsynonymous versus synonymous mutations could confirm this hypothesis.

The distribution of fitness effects **Table 2. 8** for the 1712 substitutions indicates that 67% (1141/1712) of substitutions have neutral effects on fitness, while 14% (237/1712) are slightly deleterious and 17% (288/1712) are slightly beneficial and 3% (46/1712) are beneficial.

Deep mutational scanning of an alanine racemase expressed from its chromosomal locus in *S. thermophilus*

Taking a closer look at the selected substitutions,

For substitutions resulting in stop codons, two groups of substitutions can be seen (**Figure 2. 23**), a large group of 62% (55/89) substitutions that were neutral ($E_m^r = 0.8 - 1.2$) and a second group of 26% substitutions (23/89) that are selected to a higher level ($E_m^r > 1.35$). This suggests that the expression of some truncated proteins is conferring some fitness benefit. But we do not know why this trend is not observed for all non-sense mutations.

For substitutions in the active site, there is a set of 6 substitutions that were more beneficial ($E_m^r > 1.6$) than the general trend (**Figure 2. 24**). The two mutations D170E (E_m^r 1.86) and R288K (E_m^r 1.83) in the active site entryway, these substitutions were also selected in condition C_1 . As we already hypothesised for C_1 , these mutations may be involved in decreasing a deleterious promiscuous activity or increasing a beneficial one.

Additionally substitutions S203P (E_m^r 1.72), Y263C (E_m^r 1.69), V22D (E_m^r 1.67), G220A (E_m^r 1.60) were selected. These four substitutions were purged in C_1 , and are thus clearly associated with a loss of Alr activity. Again here, we suspect that these mutations may affect promiscuous activities but without the need for maintaining Alr function. In fact, if a promiscuous activity is deleterious, there should be many mutations that could inactivate it. Hence, we hypothesize that these few mutations may be associated with a gain-of-function that is beneficial for fitness.

Results Chapter 3.2

Interestingly, Y263 is one of the catalytic residues essential for Alr activity. Previous studies in Alr^{Sp} show that mutations at Y263, completely abolish racemase activity (Y263S ~0.04 % of wild-type Alr^{Sp} racemase activity), while significantly increasing transamination rates with D-alanine (Y263S 1.7-fold higher transamination rates than wild-type Alr^{Sp}) and no transamination activity towards L-alanine (Watanabe, Yoshimura, et al. 1999). This could be a way to generate useful metabolites for the bacteria, for example, transamination could convert D-alanine and α -ketoglutarate to create pyruvate and glutamate. If true, this would be very interesting from the perspective of molecular evolution, where a new enzyme activity emerges when the environmental condition changes and the original activity becomes dispensable (availability of D-alanine).

Overall drift without selection pressure for activity (C₂) results in predominantly neutral effects of substitutions with a subset of substitutions providing beneficial effects. Although purification and extensive characterization of the mutants needs to be planned to verify the origin of the beneficial effect.

3.3. Engineering *Streptococcus thermophilus* as a reporter strain for selecting peptidic inhibitors of SARS-CoV-2 Mpro and PLpro proteases

Jason Baby Chirakadavil^{*}, Sebastian Worms^{*}, Patrice Soumillion

Lab of Biochemistry and Genetics of Microorganisms (BGM), Louvain Institute of Biomolecular Science and Technology (LIBST), Université Catholique de Louvain, Place Croix du Sud, 4-5 box L7.07.06, 1348 Louvain-La-Neuve, Belgium.

^{*} These authors contributed equally to the research work

Engineering *S. thermophilus* as a reporter strain for selecting peptidic inhibitors of SARS-CoV-2 Mpro and PLpro proteases

3.3.1. Abstract

We report the engineering of a modified alanine racemase protein in *Streptococcus thermophilus* LMD-9 strain where expression of the papain-like protease (PLpro) or the 3C-like protease (Mpro) from SARS-CoV-2 induces D-alanine auxotrophy by targeting a specifically engineered alanine racemase. Subsequently, we used this phenotype to create and screen PLpro inhibitors from four large libraries ($\sim 10^8$ transformants) encoding a total diversity of 272 million cyclic peptides. After four rounds of iterative selection and retransformation steps, we observe a marked improvement in cell viability in the cells encoding cyclic peptides. Currently, high-throughput deep-sequencing of the library is being carried out to differentiate and select true inhibitory cyclic peptides from a background of false-positives. Our results indicate that engineering of bacterial enzymes serve as a simple method to develop screens for inhibitors of viral replication. Thanks to a high transformability, the genome of *S. thermophilus* is easily modified and the combination with split intein-mediated circular ligation of peptide and proteins (SICLOPPS) is an attractive method to generate and screen millions of molecules in an inexpensive way, paving the way for cyclic-peptide drug discovery.

3.3.2. Introduction

The goal of this work was to establish an *in-vivo* system to select inhibitors of a viral protease from a library of genetically encoded

Results Chapter 3.3

randomized peptides. An *in-vivo* selection system provides multiple advantages over *in-vitro* selection or other screening methods. First, the selection occurs in the bacterial cell which is a better mimic of the complex intracellular environment of the human body, and thus improving relevance of biological activity of inhibitors. Second, there is no requirement for expensive production and purification of millions of candidate molecules to identify a target molecule (Balint and Plooy 1995; Young et al. 2011). *In vivo* methods based on a screenable phenotype like production of a chromogenic or fluorogenic compound (Cheng et al. 2007) puts a practical limit on the size of library which can be screened, even with techniques like droplet microfluidics one would need a few days of continuous operation to screen a library of 300 million variants. On the other hand, linking the selection to cell viability can allow inexpensive screening of much higher number of candidates. Here, we decided to modify an essential enzyme in *S. thermophilus*, alanine racemase (Alr) to accommodate a protease cleavage site. Since Alr activity is directly linked to cell viability in most bacteria when D-Alanine is not present in the environment (Milligan et al. 2007), targeting it serves as a perfect conduit for positive selection of inhibitors.

The proteases we chose to target are the PLpro and the Mpro proteases from the SARS-CoV-2 virus. The Mpro and PLpro proteases are involved in processing of the viral polyprotein, additionally PLpro is also involved in evading the host antiviral immune response (Bailey-Elkin et al.

Engineering *S. thermophilus* as a reporter strain for selecting peptidic inhibitors of SARS-CoV-2 Mpro and PLpro proteases

2014; Devaraj et al. 2007; Frieman et al. 2009; Shin et al. 2020; L. Zhang et al. 2020). Inhibition of either protease would prevent viral infection so both of them are attractive targets for selection of SARS-CoV-2 replication inhibitors. Pfizer is currently investigating PF-07321332, a SARS-CoV2-Mpro protease inhibitor which has demonstrated potent *in vitro* anti-viral activity against SARS-CoV-2 (currently in Phase 1b clinical trial as on March, 2021).

We chose to construct a library of cyclized peptides to screen for inhibitory molecules. Peptides provide many advantages for constructing libraries, namely the ability to be genetically encoded thus allowing the use of inexpensive combinatorial chemical synthesis for generating the DNA library and *in vivo* systems for producing the randomized peptides. The corollary of using a degenerate oligonucleotide is the ability to create very diverse libraries (20^N possibilities, where N is the length of the peptide. Moreover, peptides are amenable to head-to-tail cyclization, which is an attractive method to create small constrained molecules displaying a large density of functional groups. Constrained peptides have been shown to exhibit tighter binding than their relaxed, linear counterpart due to minimizing the loss of entropy during binding (H. Huang et al. 2021). Cyclization also significantly reduces the susceptibility to proteolysis (Kumarasinghe and Woster 2014; Satoh et al. 1996). Scott et. al. developed a method termed split intein-mediated circular ligation of peptide and proteins (SICLOPPS), taking advantage of the

Results Chapter 3.3

intramolecular splicing of a split intein (internal protein)) (Scott et al. 1999). Inteins are protein modules inserted in the sequence of another protein. After translation of the host protein, the intein catalyzes its self splicing from the host, leaving behind an intact sequence. Split inteins are made of two fragments, that can be fused, in a permuted order, to the extremities of a polypeptide chain. After translation, the two intein fragments form an active intein that catalyse the circularization of the target sequence. Early applications of SICLOPPS used the split intein responsible for the assembly of the DnaE replicative polymerase of *Synechocystis* sp. PCC6803 (Scott et al., 1999). However, analysis of related inteins showed that the intein from the DnaE of *Nostoc punctiforme* had much higher activity (Zettler, Schütz, and Mootz 2009). Experience in our lab showed that the accumulation of spliced intein fragments can be toxic to the cell. To minimize this issue, we used the Npu-ssrA construct reported by Townend & Tavassoli. There, the C-terminal fragment is fused to the ssrA degradation tag and recycled by the ClpX protease (Townend and Tavassoli 2016) in *E. coli* and by the ClpC protease in *S. thermophilus*. SICLOPPS has been previously used to select inhibitors of viral proteases, protein-protein binding, methyltransferases and many other proteic interactions (Block and Grafstrom 1990; Cheng et al. 2007; Naumann, Tavassoli, and Benkovic 2008; Tavassoli et al. 2008; Young et al. 2011).

Engineering *S. thermophilus* as a reporter strain for selecting peptidic inhibitors of SARS-CoV-2 Mpro and PLpro proteases

Here we report the construction of a screening strain of *Streptococcus thermophilus* where growth defect on defined growth medium is linked to SARS-CoV-2 Mpro and PLpro protease activity. This was achieved by modifying the native alanine racemase so that the cells become auxotrophic to D-alanine on expression of the protease. We demonstrate the utility of this system by generating and screening an *in-vivo* library of more than 2.72×10^8 peptides. Currently, deep-sequencing analysis is being carried out to precisely identify candidate inhibitors.

3.3.3. Results

Engineering AlrSt insertants that are cleaved by SARS-CoV2 protease

The first step was to identify a site for inserting the protease recognition sequence in AlrSt (Alr from *S. thermophilus*) without disrupting its activity. For this, we searched multiple sequence alignments for indels occurring naturally in homologous Alr sequences. We observed that AlrSt position 192 accommodated inserts in homologous Alr sequences from other species. Analysis of the structure revealed that positions T192, D193, R194 are situated in a loop far from the active site or dimer binding site. Based on these, we decided to create small Alr library with a semi-rationally designed insert comprising the protease cleavage site (*Figure 3. 1 A*). A partially degenerate TDR motif was duplicated on the flanking sides of the insert, allowing the native residue at each position as well as glycine

Results Chapter 3.3

and eventually two other residues depending on the genetic code. The protease cleavage site was placed between these motifs and a fully degenerate codon was added at both junctions .

Construction of the *alr* carried out by replacing the *alr* gene with the degenerated *alr* insert library by the two step *cat-oroP* scarless gene replacement strategy (Figure 3. 1 B). Degenerate oligonucleotides were obtained as primers, while the wild-type AlrST sequence was used as a template for PCR amplification followed by Gibson assembly to assemble the fragments. This product was used as the donor DNA for natural transformation. We performed a simultaneous selection for loss of the *oroP* gene and alanine racemase activity by plating the cells on media devoid of D-alanine, supplemented by 1.5 mg/ml 5-FOA. Out of 10⁴ transformed cells 3.28% encoded an active Alr while accommodating the Mpro site, and 0.35% encoded an active Alr accommodating the PLpro site (Table 3. 1).

Table 3. 1. Size of *alr*^{insert} library and the percentage of active Alr^{insert} colonies.

<i>alr</i> ^{insert}	Total transformants (cfu)	Alr ^{insert} active (cfu)	Percentage Alr ^{insert} active
<i>alr</i> ^{Mpro}	1.08 x 10 ⁴	330	3.28%
<i>alr</i> ^{PLpro}	1.5 x 10 ⁴	54	0.35%

Engineering *S. thermophilus* as a reporter strain for selecting peptidic inhibitors of SARS-CoV-2 Mpro and PLpro proteases

Active $\text{Alr}^{\text{insert}}$ are the colonies able to grow on growth media without D-alanine. *alr* encoding the Mpro protease recognition site is denoted as alr^{Mpro} for the gene or Alr^{Mpro} for the protein, *alr* encoding the PLpro recognition site is denoted as $\text{alr}^{\text{PLpro}}$ for the gene or $\text{Alr}^{\text{PLpro}}$ for the protein.

Sequencing some of the colonies and eliminating duplicates yielded 6 unique sequences encoding Alr^{Mpro} and 6 unique sequences encoding $\text{Alr}^{\text{PLpro}}$. All the 12 variants demonstrated wild type like D-alanine prototrophy and accommodated the full sequence of the recognition site.

Results Chapter 3.3

A

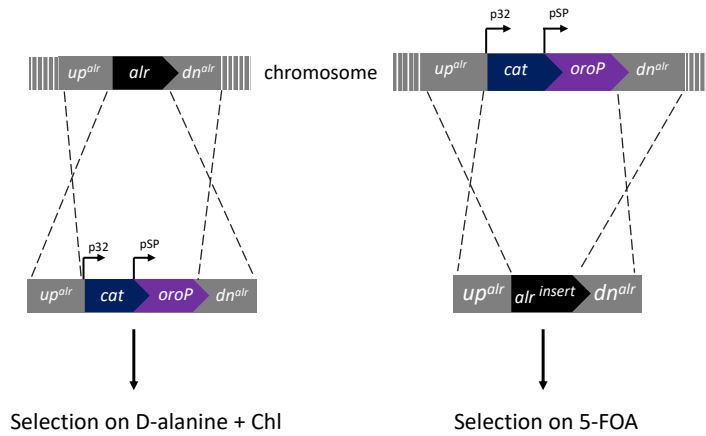
alr^{Mpro} insert library

L191-[T/G/S/A][D/G][R/G]XSAVLQ↓SGFRKX[T/G/S/A][D/G][R/G]-P195(P210)

alr^{PLpro} insert library

L191-[T/G/S/A][D/G][R/G]XRELNGG↓AX[T/G/S/A][D/G][R/G]-P195(P210)

B



C

$\Delta alr::alr^{Mpro}$ insert sequences

L191-GGGSSAVLQ↓SGFRKNTGR-P195(P210)
L191-GGRSSAVLQ↓SGFRKSGR-P195(P210)
L191-GDRHSAVLQ↓SGFRKSAGR-P195(P210)
L191-GGGA SAVLQ↓SGFRKSGDR-P195(P210)
L191-ADRRSAVLQ↓SGFRKASGG-P195(P210)
L191-GGRSSAVLQ↓SGFRKSGR-P195(P210)

$\Delta alr::alr^{PLpro}$ insert sequences

L191-SGRQRELNGG↓AMSDG-P195(P210)
L191-SDRYRELNGG↓AMSGR-P195(P210)
L191-SDRNRELNGG↓ANS DR-P195(P210)
L191-SGRPRELNGG↓ASGGR-P195(P210)
L191-GDRCRELNGG↓AISGR-P195(P210)
L191-SDRARELNGG↓APTDR-P195(P210)

Engineering *S. thermophilus* as a reporter strain for selecting peptidic inhibitors of SARS-CoV-2 Mpro and PLpro proteases

Figure 3. 1. Directed evolution of the TDR loop of *alr*ST to accommodate the insertion of recognition sequence of SARS CoV-2 Mpro and PLpro proteases. (A) Design of the semi degenerated codon library flanking the insertion sequence between Leucine 191 and Proline 195. Numbers in brackets denote the new position of the amino acid with the insert. **(B)** Schematic representation of the two step *cat-oroP* based scarless replacement of the wild-type *alr* with the *alr* degenerated codon library. **(C)** Insertion sequences for selected colonies encoding an active Alr and the recognition sequence of the protease.

Cloning Mpro or PLpro in the *alr*^{insert} background generates D-alanine auxotrophy

Since all 6 selected insertants of *alr*^{Mpro} and *alr*^{PLpro} displayed similar D-alanine prototrophy phenotype, so we used a mix of all 6 insertants for each strain, for the transformation steps.

Mpro or PLpro protease genes were introduced at the *tRNA_{serine}* locus under an inducible xylose promoter, by the two step *cat-oroP* scarless gene replacement strategy (**Figure 3. 2 A**). We chose *tRNA_{serine}* locus for its permissiveness to insertions (Mignolet et al. 2018). 49 colonies were subsequently picked and screened for lack of growth on media supplemented with xylose and wildtype like growth in media supplemented with D-alanine. Out of all the different screened colonies, we choose the colony with the best growth in media supplemented with D-alanine and the least growth in media supplemented with xylose.

Results Chapter 3.3

Colonies demonstrating the expected phenotype were sequenced at the *alr* locus to reveal the sequence of the insert **ADRR**SAVLQSGFRK**ASGG** in the strain $\Delta alr::alr^{Mpro}$ P_{xyl2} -*mpro* and **SDR**ARELNGGAP**TDR** in $\Delta alr::alr^{PLpro}$ P_{xyl2} -*plpro*.

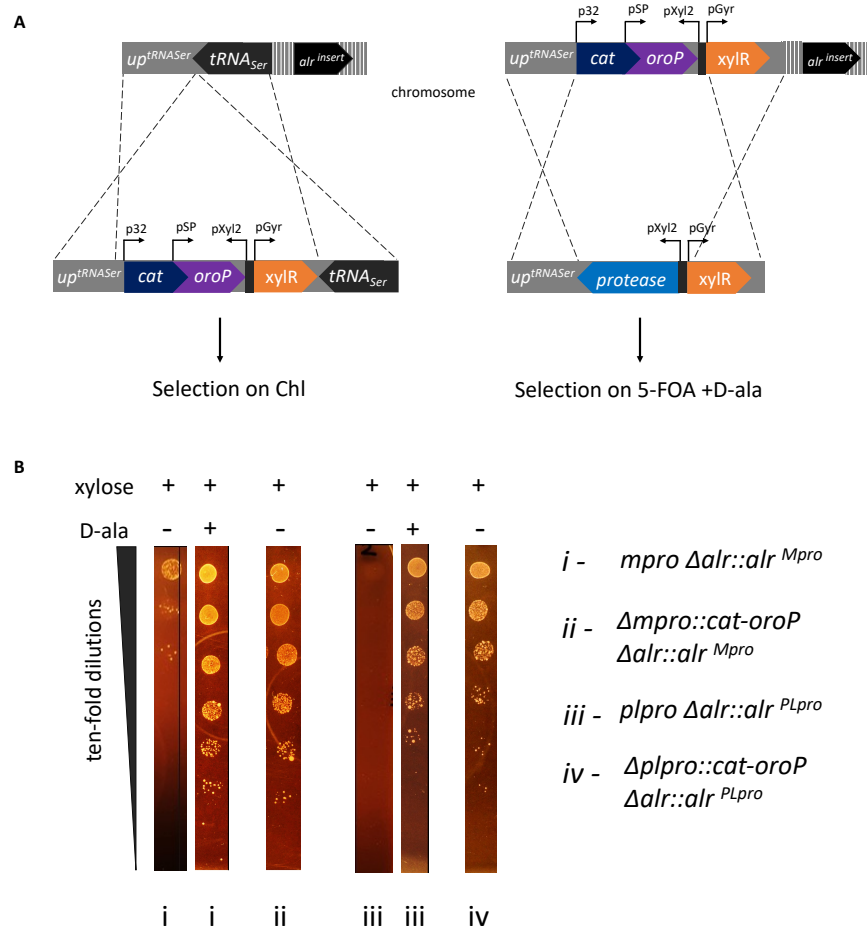


Figure 3. 2. Construction of PLpro induced D-alanine auxotrophic strain.

(A) Schematic representation of construction of *protease* $\Delta alr::alr^{insert}$

Engineering *S. thermophilus* as a reporter strain for selecting peptidic inhibitors of SARS-CoV-2 Mpro and PLpro proteases

strains **(B)** Constructed mutant strains show D-alanine auxotrophy phenotype conferred due to the protease cleavage of the modified Alr protein encoded in the strain. Deletion of the protease gene restores D-alanine prototrophy.

We then verified that the deletion of the protease (by gene replacement with $P_{32-cat} P_{Sp-oroP}$) in both strains results in a full reversal of the induced D-alanine auxotrophy phenotype, further confirming that the D-alanine auxotrophy is the result of protease cleavage of Alr^{Mpro} and Alr^{PLpro} (Figure 3. 2 B). Spotting ten-fold dilutions on growth medium devoid of D-alanine reveals revertant colonies where the induced D-alanine auxotrophy phenotype is lost. When individual revertant colonies were picked, regrown and respotting with dilutions, similar number of colonies were counted with and without D-alanine, suggesting that these are true genetic revertants. Sequencing some of the revertants revealed mutations in either the protease gene or the *alr* sequence. Interestingly, mutation R208L (R196L) also caused a full reversal of the auxotrophy phenotype in the *alr*^{PLpro} gene, however it is not clear how this mutation interferes with the protease activity since it is neither situated on the protease recognition site nor the TDR motif. Further study is needed to identify the exact mechanism of reversion for some of these mutants.

The next step was to use these strains as screening strains to select inhibitors of the proteases by expressing a library of molecules from

Results Chapter 3.3

within the cell and selecting cells for D-alanine prototrophy. Since the background frequency of revertant emergence for *mpro Δalr::alr^{Mpro}* strain was relatively high (1 in 3500 cfu), while the revertant emergence for *plpro Δalr::alr^{PLpro}* (1 in 100000 cfu) was acceptable for screening. Hence, we decided to go ahead with only the *plpro Δalr::alr^{PLpro}* strain for further screening.

Tentative selection of cyclic peptide inhibitors of SARS-CoV-2 PLpro

Next, we used *plpro Δalr::alr^{PLpro}* strain to search for PLpro inhibitory cyclic peptides encoded *in-vivo* using split intein-mediated circular ligation of peptide and proteins (SICLOPPS) technology (Scott et al. 1999). Previously, similar approaches have been successfully utilized to select protease inhibitors, either based on a selectable phenotype (Dautin et al. 2000) or linked to cell viability (Balint and Plooy 1995; Block and Grafstrom 1990; McCall, Kadam, and Katz 1994). We designed four SICLOPPS peptide libraries using oligonucleotide primers with degenerated codons. We used a similar strategy for the construction of the first three libraries, a cysteine followed by three, five or seven randomised codons, creating the libraries with the motif CX₃, CX₅ and CX₇ (X = NNK, where N is A, T, C, or G; and K is G or T). For the fourth library, we included the PLpro recognition site (LXGG) in the design through a semi degenerated codon library, creating the motif CX₃LZGGX₂ (Z is either L, K, N or S, encoded by ARV). ARV degeneracy was used for the 6th position because Arginine, Lysine and Asparagine are the commonly observed amino acids in SARS-CoV-2 PLpro

Engineering *S. thermophilus* as a reporter strain for selecting peptidic inhibitors of SARS-CoV-2 Mpro and PLpro proteases

cleavage sites. Cysteine was fixed in all the libraries since it is essential for intein-mediated peptide cyclization (Scott et al. 1999).

We decided to clone the peptide library under the control of the *leuB* promoter. Indeed, we previously demonstrated that the transcriptional activity of this promoter is higher compared to P_{xyI2} and we need more peptide than protease in the cell. For a markerless insertion of the cyclic peptide library in the chromosome, replaced the *leuB* gene by libraries of cyclic peptide encoding genes. The donor DNA for the cyclic peptide library was prepared by Gibson assembly and used for natural transformation of the recipient strain where the P_{32-cat} $P_{SP-oroP}$ cassette was replaced by the library and transformants were selected on solid growth media supplemented with 5-FOA.

Results Chapter 3.3

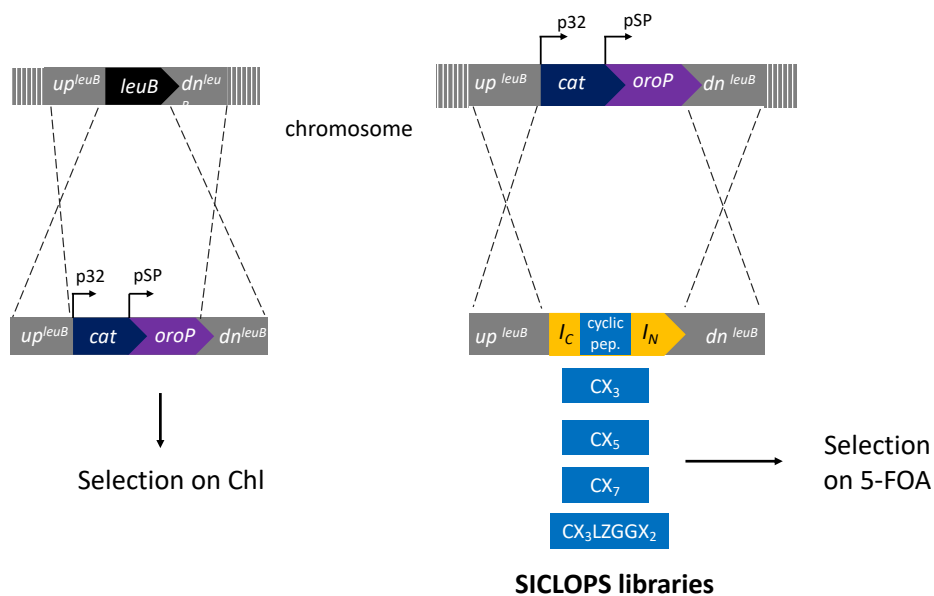


Figure 3. 3. Design and construction of the cyclic peptide libraries. A cassette consisting of *P₃₂-cat P_{SP}-oroP P_{Xyl2} P_{gyr-xyIR}* was inserted replacing the wild-type *leuB* gene, transformants were selected for gain of chloramphenicol resistance. Subsequently, the gene encoding the cyclic peptide library was introduced under the control of the *leuB* promoter (part of the leucine operon), transformants were selected for gain of 5-FOA resistance. Independent transformations were carried out for each of the four libraries.

In each of the libraries we obtained more than 10^8 transformants, with this size of library we can expect 100% coverage of the theoretical diversity of *CY₃*, *CY₅* and *CY₃LZGGY₂* libraries (**Table 3. 2**). For the *CY₇*

Engineering *S. thermophilus* as a reporter strain for selecting peptidic inhibitors of SARS-CoV-2 Mpro and PLpro proteases

library, we can expect around 21% of the theoretical diversity to be covered, as calculated by the GLUE-IT algorithm (Firth and Patrick 2005; Patrick, Firth, and Blackburn 2003).

Table 3. 2 The randomization scheme of the four libraries.

Library name	Randomisation scheme	Theoretical diversity	Total transformants	Expected coverage
CX ₃	- Cys-(NNK) ₃ -	8×10^3	1.33×10^8	100 %
CX ₅	- Cys-(NNK) ₅ -	3.2×10^6	3×10^8	100 %
CX ₇	- Cys-(NNK) ₇ -	1.28×10^9	3×10^8	20.89 %
CX ₃ LZGGX ₂	- Cys-(NNK) ₃ -Leu-(ARV)-Gly-Gly-(NNK) ₂ -	1.28×10^7	1.33×10^8	100 %

Here, X stands for the NNK degenerate codon, while Z stands for the ARV degenerate codon. The expected coverage prediction was calculated by using the GLUE-IT algorithm (Firth and Patrick 2005; Patrick, Firth, and Blackburn 2003)

As an additional verification of the construction of the library we randomly picked and sequenced colonies from the library to calculate the percentage of colonies encoding the full construct without indels **Table 3. 3**. For the CY₃ library, 90.47% of colonies sequenced were properly constructed. For the CY₅ library, 88.46% of colonies sequenced were properly constructed. For the CY₇ library, 80% of colonies sequenced were properly constructed. For the CY₃LZGGY₂ library, 65% of colonies sequenced were properly constructed.

Results Chapter 3.3

Table 3. 3. Percentage of libraries properly constructed without indels.

Library	Colonies sequenced	Percentage of colonies built correctly	
		n	%
CX₃	21	19	90.47 %
CX₅	26	23	88.46 %
CX₇	20	16	80 %
CX₃ LZGGX₂	20	13	65 %

n represents the number of colonies sequenced.

Only cells in the library encoding a protease inhibiting cyclic peptide should be protected from protease cleavage of Alr^{PLpro} and thus should demonstrate D-alanine prototrophy and the ability to grow on solid growth medium devoid of D-alanine. To select, isolate and identify inhibitors, we first grew the libraries on large petri plates with solid growth medium supplemented with 0.25 % xylose devoid of D-alanine. After the first round of selection, we observed no significant change in the percentage of D-alanine prototrophic mutants (0.003 %) as compared to the control *alr^{PLpro} plpro* strain. This suggested that there is a high frequency of spontaneous D-alanine prototroph mutants due to mutations unrelated to peptide inhibition. Simultaneous sequencing of

Engineering *S. thermophilus* as a reporter strain for selecting peptidic inhibitors of SARS-CoV-2 Mpro and PLpro proteases

the three loci- the 5.853 kb *plpro* gene, the recognition sequence in *alr*^{PLpro} and the encoded cyclic peptide revealed that 'false positive' D-alanine prototrophy can result from deleterious mutations in the *plpro* gene or mutations in *alr*^{PLpro} modifying the PLpro recognition site. We surmised that multiple rounds of selection followed by retransformation into the fresh 'parent' strain *alr*^{PLpro} *plpro* Δ *leuB::P₃₂-cat* *P_{SP}-oroP* would allow enrichment of the frequency of true positive cyclic peptide conferred prototrophs above the background frequency of prototrophy due to spontaneous mutations.

Results Chapter 3.3

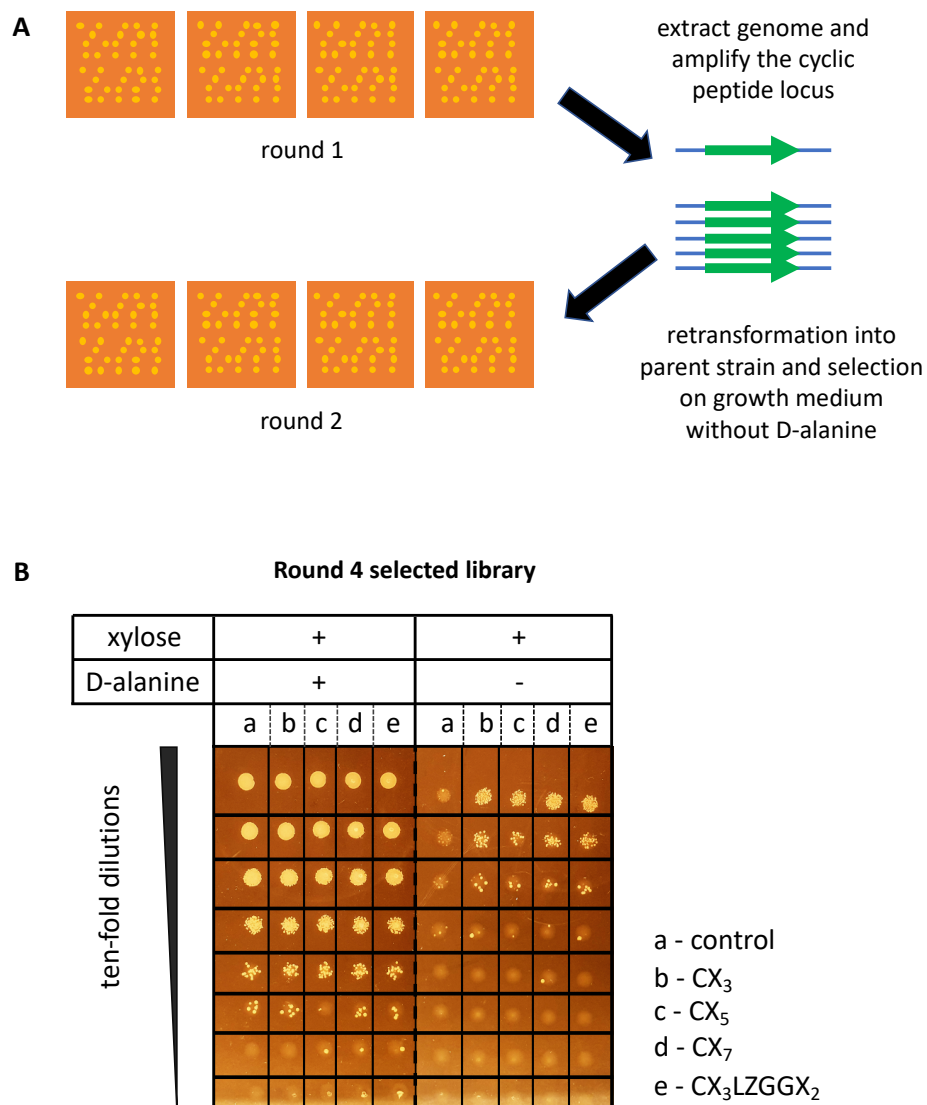


Figure 3. 4. Enrichment of PLpro inhibitory cyclic peptides (A) Scheme illustrating by successive rounds of selection of the library and retransformation of the parent strain. The library was grown on large plates containing solid growth media supplemented with 0.25 % xylose,

Engineering *S. thermophilus* as a reporter strain for selecting peptidic inhibitors of SARS-CoV-2 Mpro and PLpro proteases

devoid of D-alanine to select cyclic peptides which can inhibit PLpro mediated cleavage of Alr^{PLpro}. Next the colonies which grew were scraped and the genome was extracted and used as template for PCR amplification of the cyclic peptide library. This PCR product was used as the donor DNA for natural transformation of the parent strain. This process was repeated for each round of selection. **(B)** Comparison of the D-alanine auxotrophy of the round 4 selected libraries. For the control strain transformation, the *cat-oroP* cassette at the *leuB* locus was replaced with the wild-type *leuB* gene; while for each of the libraries, the *cat-oroP* cassette at the *leuB* locus was replaced with the PCR amplified donor DNA from the round 3 selected cyclic peptide library.

For retransformation of the parent strain, we scraped cells from round 1 of selection, lysed and extracted their genomic DNA, subsequently 2 µg of the genomic DNA was used as template for PCR amplification of the locus corresponding to the cyclic peptide library, 1.5 kb flanking regions were additionally included to function as homologous arms for recombination, this amplicon was directly used as donor DNA for natural transformation of the parent strain on growth media with 2mg/ml 5-FOA without supplementation with D-alanine. We iterated these steps for round 2, round 3 and round 4 of selection. Ultimately, we plated dilutions of the round 4 selected cells on growth medium with and without D-alanine to count the proportion of the library demonstrating D-alanine prototrophy.

Results Chapter 3.3

We observed similar number of colonies after 24 hrs for the control *alr*^{PLpro} *plpro* strain $\Delta P_{32-cat} P_{SP-oroP} :: leuB$ as compared to the transformants encoding the library, although clearly the round 4 cyclic peptide library transformants demonstrated much larger colony sizes. This led us to hypothesise that the cyclic peptides were conferring a growth advantage. Out of 9 colonies sequenced from the CX₅ library, the peptide CAAAAA, CTTNAS and CVLATA appeared in both round 1 and 4. Sequencing of CX₃, CX₇ and CX₃LZGGX₂ libraries failed to present a clear consensus motif for the cyclic peptides. Again, some of these colonies were observed to have mutations in the *plpro* gene while some others had mutations in the *alr*^{PLpro} recognition sequence. As expected, some of these colonies also encoded cyclic peptide sequences with stop codons or mutations in the c- or n-intein fragment, suggesting that even in the round 4 library, there is a mix of true positive peptides and false positives. Due to the practical constraint for simultaneously monitoring spontaneous mutations at three loci- the 5.853 kb *plpro* gene, the recognition sequence in *alr*^{PLpro} and the encoded cyclic peptide, to distinguish false positives from true inhibitory cyclic peptides; we decided to sequence the genetic diversity of the selected round 1 and round 4 libraries using next generation sequencing. Sequencing results are pending and, because of time constraints, we had to stop working on this project.

1.3.1. Discussion

Engineering *S. thermophilus* as a reporter strain for selecting peptidic inhibitors of SARS-CoV-2 Mpro and PLpro proteases

Here we report an *in-vitro* engineered bacterial system where expression of a heterologous protease confers an auxotrophy phenotype which can be used for screening inhibitors of the protease. In a study, Ast et. al. have shown that expression of HIV-1 protease is toxic to *Escherichia coli* cells and they used this system to select protease inhibitory peptides (Ast et al. 1998). But unless expression of an enzyme induces cell death via toxicity this method cannot be directly employed for development of inhibitors for other viral proteins. Previous other groups have developed systems for artificially linking bacterial viability to presence of a protease and subsequently using the system to screen inhibitors (Balint and Plooy 1995; Block and Grafstrom 1990; Naumann, Tavassoli, and Benkovic 2008). Balint and Plooy developed an inactive fusion of the S12 ribosomal protein with the Zucchini yellow mosaic virus protease recognition sequence, such that cleavage of the recognition sequence confers streptomycin sensitivity. But this system had a high background revertant frequency and the screening of thousands of candidates was necessary to select true positive phenotypes. Similar to our approach, Block & Grafstrom modified a tetracycline gene by inserting a human immunodeficiency virus (HIV) protease recognition sequence. Coexpression of the HIV protease with the tetracycline gene prevented growth of the cells in presence of tetracycline. The strategy in the present work additionally builds upon these approaches. However, while other systems relied on introduction of heterologous genes, either to induce resistance in the work of Block &

Results Chapter 3.3

Grafstrom, or interact with cellular proteins owing to a dominant phenotype in the work of Balint and Plooy; we on the other hand modified the native alanine racemase gene to create a target for protease cleavage, thus creating a selectable phenotype. Secondly, while previous studies cloned the modified enzymes on heterologous plasmid vectors, we have cloned the modified *alr* gene, inserted the protease gene and encoded a inhibitor library, all on the chromosome of *S. thermophilus*. Chromosomal gene expression allows for more stable and uniform expression of the genes as compared to plasmid expression (Ryan et al. 2014).

Engineering *S. thermophilus* as a reporter strain for selecting peptidic inhibitors of SARS-CoV-2 Mpro and PLpro proteases

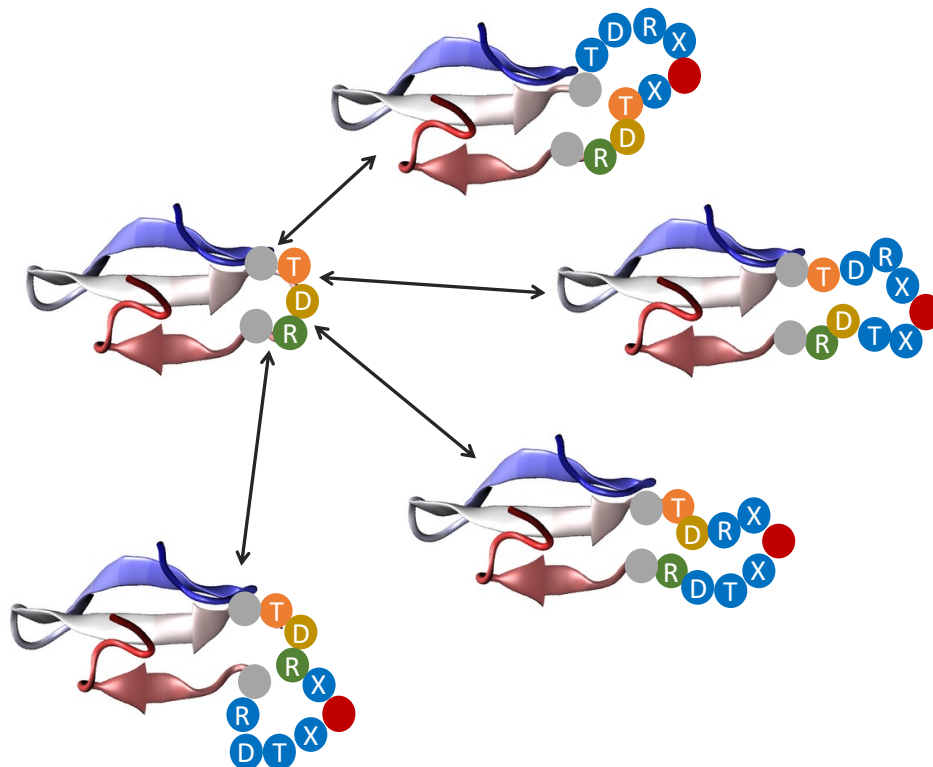


Figure 3. 5. A strategy for facilitating insertion tolerance: duplication of the insertion site (TDR loop in our case) allows multiple insertional topologies where residues before and/or after the insert may occupy wild-type structural positions. Spheres in orange, yellow and green represent the residues positioned in the wild-type configuration, grey spheres represent the residue adjacent to the motif, blue spheres represent the residues in the duplicated loop positioned in a non-wild-type configuration, red sphere represents the insert sequence. Black arrows point to the sites where the insert can be accommodated while still allowing for wild-type-like configuration for the TDR residues. The

Results Chapter 3.3

degeneracy allows for wild-type and glycine residues at the original and duplicated TDR motif for allowing loop flexibility, and additionally full NNK degeneracy for one residue flanking the recognition site (X = NNK in the figure).

Additionally, we describe a simple method to accommodate inserts without disrupting enzyme activity. This method involves identification of a position or motif based on indel detection in phylogeny and the design of a small degenerate library where insertion is combined with a small duplication of the site (**Table 3. 3**). We also describe how this conditional auxotrophy system can be used to screen a massive SICLOPPS library of 272 million *in-vivo* encoded molecules, taking advantage of the highly efficient natural competence of *S. thermophilus*. This approach was similar to the work by Young et. al. from Peter Schultz's lab, additionally they also include an expanded set of amino acid building blocks by use of orthogonal aminoacyl-tRNA synthetase/tRNA_{CUA} pairs (Young et al. 2011). This could be an interesting addition to our system. However, albeit employing a similar selection strategy, they do not report issues in selection due to high spontaneous frequency of false positives. It is known that expression of viral proteases can exhibit toxicity in bacteria (Ast et al. 1998; Babél, Linneversl, and Schmidt 2000). High frequency of mutation suggests that the PLpro / Mpro proteases induce toxicity due to leaky expression from the pXyl2 promoter. Perhaps, a tighter control of transcription by knocking down of gene expression by CRISPR interference

Engineering *S. thermophilus* as a reporter strain for selecting peptidic inhibitors of SARS-CoV-2 Mpro and PLpro proteases

(Kjos 2019) can keep PLpro / Mpro levels low enough that spontaneous mutants do not accumulate. The other alternative strategy is to employ next generation sequencing as explained below.

Future work

Next generation sequencing could be employed to distinguish true positive cyclic peptide candidates from false positives

Since sequencing individual colonies failed to provide the statistical power to correctly distinguish the false positives from the true inhibitory cyclic peptides, we decided to opt for the greater statistical power of next generation sequencing. Due to degeneracy of genetic code a single cyclic peptide sequence can be encoded by several distinct nucleotide sequences. Thus, if the abundance of a cyclic peptide sequence is based on the existence of spontaneous mutations elsewhere in the genome, they would predominantly be encoded by the same nucleotide sequence. Presence of different nucleotide sequences in the cyclic peptide library encoding the same amino acid motif would be a clear indication of selection for PLpro protease inhibition. We have verified that *in-vitro* transcription translation (IVTT) systems can be used for production of cyclic peptides affording a potential screening approach. Currently, we are working on the next generation sequencing of the selected libraries to create a subset of candidate cyclic peptides. Later these cyclic peptides

Results Chapter 3.3

will be produced *in-vitro* and the inhibition will be verified using a PLpro fluorogenic assay.

4.

Materials and Methods

In this chapter we explain the materials and methods for the experimentation. The procedures for making chromosomal edits and library preparation in *S. thermophilus* and *E.coli* was published by us as a book chapter (me, as co-first author). Part of the chapter which deals with protocols in *S. thermophilus* was completely developed/written by me, edited by Patrice Soumillon and Pascal Hols. In the first part of this chapter, I will directly attach (with permissions from Springer Nature), part of the published book chapter with minor adaptation.

Reprinted/adapted by permission from Springer Nature: Building Scarless Gene Libraries in the Chromosome of Bacteria by Gol Mohammad Dorrazehi, Sebastian Worms, Jason Baby Chirakadavil et al [License Number : 5081000827504] (2020)

Building Scarless Gene Libraries in the Chromosome of Bacteria

4.1. Building Scarless Gene Libraries in the Chromosome of Bacteria (Modified from (Dorrazehi et al. 2020))

4.1.1. Growth media

S. thermophilus strain LMD-9 was purchased from ATCC (BAA-491). *S. thermophilus* cells were grown at 37°C in an anaerobic jar (BBL GasPak Systems; Becton Dickinson, Franklin Lakes, NJ) for solid media, and in eppendorf (without shaking) for liquid culture in skimmed milk medium (SMM, Campina, Belgium), M17 medium (Difco) or chemically defined medium (CDM) as described in Letort & Julliard, 2001 (Letort and Juillard 2001). M17 and CDM were supplemented with 1% (w/v) lactose (M17L and CDML, respectively). Supplements were added to the different media at the following concentrations: chloramphenicol (5 mg/L), erythromycin (2.5 mg/L), 5-fluoroorotic acid (5-FOA) (500 mg/L unless specified otherwise) (Melford, UK). 5-FOA was diluted in DMSO at a concentration of 100 mg/mL. Chemicals were purchased from Sigma-Aldrich, unless otherwise specified (Fontaine et al. 2013b; Gardan et al. 2013b).

The *S. thermophilus* genome sequence can be accessed from GenBank with accession no. CP000419 (LMD-9).

Here is a list of genes used in this work with their annotations.

Gene names	Annotation
<i>alr</i>	STER_RS08315

Materials and Methods Chapter 3.3

<i>leuB</i>	STER_RS05775
<i>lysA</i>	STER_RS01450
<i>tRNAserine</i>	STER_RS07710

4.1.2. Natural transformation

All transformations were performed in CDML or SMM, the later affording higher transformation rates. The synthetic XIP peptide (NH₂-LPYFAGCL-COOH) was ordered at Peptide 2.0 (Chantilly, VA), resuspended at a concentration of 500 µM in 100% DMSO, and stored at -20°C in 20 µL aliquots.

4.1.3. Oligonucleotides

All primers were purchased from Eurogentec (Seraing, Belgium) The sequences in bold and black are the 3'-end which hybridize to the target, while the red sequences are the 5' overhang regions necessary for fragment assembly. The sequence in italic and blue for primer. Annealing temperatures (T_m) were predicted taking only the hybridizing sequences of the primers using the online tool available at <http://tmcalculator.neb.com/>.

PCR/Target to amplify	Primer Pair	Sequence (5'→3')
-----------------------	-------------	------------------

Building Scarless Gene Libraries in the Chromosome of Bacteria

U-tRNA _{ser} Rec	K1	F w	AACTATCATCCGTCATCC
		Rv	CCTTATGGGATTTATCTTCCTTACAGAAATAGACAAAGTGTC
D-tRNA _{ser} Rec	K2	F w	CCGGAATAATATTCATTAAATTTTATAAAATT TGTCTATTTCTGAAAAGG
		Rv	GATGGTCGTATTTTGACC
P32-cat	K3	F w	GGACACTTTGTCTATTTCTGTAAGGAAGATAA ATCCCATAGG
		Rv	GCAAATGTACTGTCAAGAATAATTCACGTTACTA AAGGGAATG
PSP-oroP	K4	F w	TACATTCCTTTAGTAACGTGAATATTCTTGAC AGTACATTTGC
		Rv	CCTTTTCAGAAATAGACAAATT TTATAAAAATT TAATGAATATTATTCCGG
U-tRNA _{ser} Lib	K5	F w	GGCTTTTCAAACTATCATCCGTCATCC
		Rv	GGTTCCTCGGGTAGCGGA
D-tRNA _{ser} Lib	K6	F w	CATAATTTGTCTATTTCTGAAAAGGAGAGGAG GGG
		Rv	GTCTGGGAAATGGTGATATGATGAGTGC
leuB-Lib	K7	F w	TCCGCTACCCGAGGAACC
		Rv	CCCCTCCTCTCCTTTTCAGAAATAGACAAATTATG

Table 4. 1 List of primers used in this book chapter

4.1.4. Transformation in *S. thermophilus* using natural competence

Principle

Natural transformation is the capability of several species of bacteria to achieve a transient competent state in which they are able to take up exogenous DNA inside the cells. This DNA can be used by the cells to fulfill nutritional requirements for deoxyribonucleotides, or depending on the sequence similarity to the chromosome, it can be integrated into the chromosome by homologous recombination (Mell and Redfield 2014). *S. thermophilus* is one of the species in which natural transformation has been well described (Blomqvist, Steinmoen, and Håvarstein 2006). The bacterium is not pathogenic (GRAS status, Generally Recognized As Safe). Some strains featuring very high transformation efficiencies have been reported and, combined with a highly efficient intrinsic recombinogenic activity, those strains are interesting candidates for the creation of massive chromosomal libraries (Fontaine, Dandoy, Boutry, Delplace, de Frahan, et al. 2010). In our hands, the method described herein has been successfully used for transforming up to 50% of the whole bacterial population, indicating that libraries containing up to ten billion individual transformants can be easily built.

In *S. thermophilus*, under exponential growth conditions, the expression of *comS* is triggered, leading to the production of the precursor (ComS) of the peptide pheromone. Then, ComS is exported and matured extracellularly by a protease cleavage to produce mature ComS, also called XIP (*comX* Inducing Peptide). Above a certain concentration

Building Scarless Gene Libraries in the Chromosome of Bacteria

threshold, extracellular XIP initiates peptide (re-)importation by the Ami/Opp transporter into the cytoplasm. Intracellularly, it interacts with the transcriptional activator ComR, forming a dimeric ComR-XIP complex that binds to the ComR box located in the promoter of *comS* and *comX* genes to promote their transcription, and thus inducing a positive feedback loop through the induction of *comS* (see **Figure 4. 1**). Eventually, ComX induces transcription of late competence genes necessary for transcription of genes involved in DNA uptake and processing (Blomqvist, Steinmoen, and Håvarstein 2006; Fontaine, Boutry, Henry De Frahan, Delplace, Fremaux, et al. 2010; Fontaine et al. 2013b; Haustenne et al. 2015). Competence can be artificially induced by adding a synthetic XIP peptide (ComS₁₇₋₂₄ LPYFAGCL) to the growth medium (Blomqvist, Steinmoen, and Håvarstein 2006; Fontaine, Dandoy, Boutry, Delplace, de Frahan, et al. 2010).

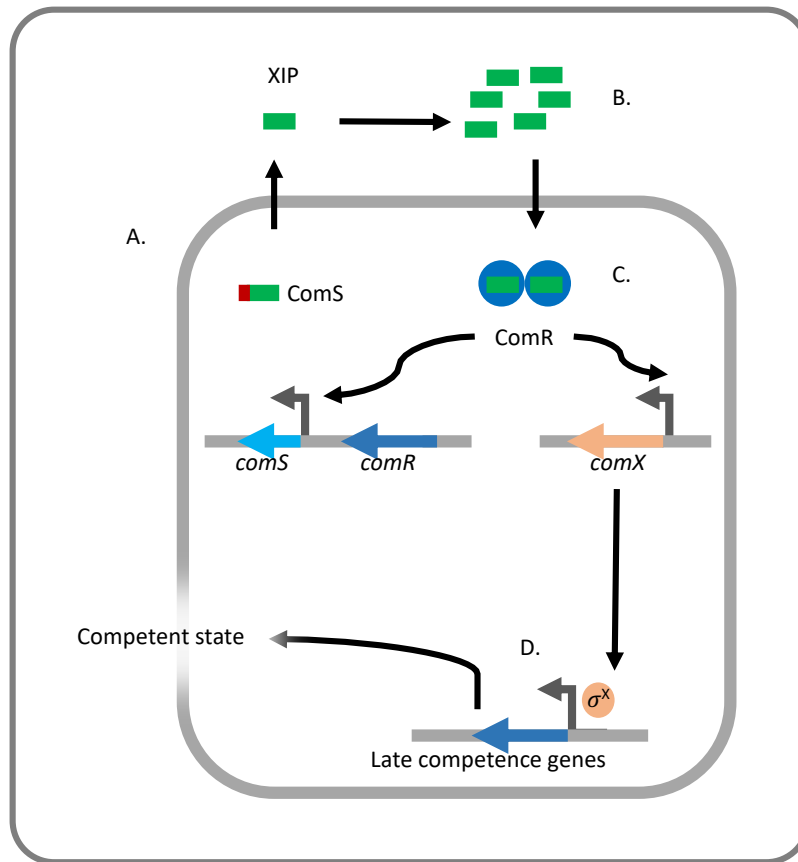


Figure 4. 1 Competence regulation by the *comRS* system in *S. thermophilus*. (A) Expression of *comS* initiates the system; the precursor peptide pheromone ComS is exported and matured into XIP. (B) When reaching a threshold extracellular concentration, the XIP pheromone is reimported into the cell. (C) In the cytoplasm, XIP interacts with ComR. The dimeric ComR-XIP complex activates transcription of genes *comS* and *comX* (D) ComX induces transcription of late competence genes required for DNA transformation. Figure modified from (Haustenne et al., 2015).

Building Scarless Gene Libraries in the Chromosome of Bacteria

Naturally competent bacteria offer a powerful approach to directly manipulate the chromosome of bacteria. Typical methods to modify the genome involve creating bicistronic cassettes including antibiotic resistance gene in addition to the gene of interest. However, this can be complicated when multiple mutations have to be introduced at different genomic loci, increasing the number of resistance gene required (Sung et al. 2001). An elegant scarless strategy to circumvent the problem is based on the replacement of a negative selection marker by the gene or the library of interest.

Our method (outlined in **Figure 4. 2**) was developed with the orotate transporter (*oroP*)-based selection/counter selection marker, as described previously (Solem et al. 2008). The presence of the *oroP* gene confers the ability to utilize orotate as the sole pyrimidine source in a pyrimidine auxotrophic strain, therefore acting as a positive selection marker. Uracil auxotrophic mutants could be selected by the method described previously (Martinussen et al. 2001). Additionally, the presence of *oroP* results in sensitivity to 5-FOA. In our method, we only utilized this property by selecting for 5-FOA resistance as a consequence of *oroP* replacement by the library insertion. Before library construction, *S. thermophilus* is therefore modified by inserting a bicistronic cassette carrying *oroP* and the *cat* gene in the target locus. The *cat* gene encodes chloramphenicol acetyl transferase, it is expressed from the P₃₂ constitutive promoter and confers resistance to chloramphenicol (Cm^R)

Materials and Methods Chapter 3.3

while *oroP* is expressed under a synthetic promoter as described previously (Defoor, Kryger, and Martinussen 2007). Another system for negative selection is based on the *rpsL* gene (Sung et al. 2001). In our experience however, the *rpsL* system generates a high frequency of false-positive transformants.

Building Scarless Gene Libraries in the Chromosome of Bacteria

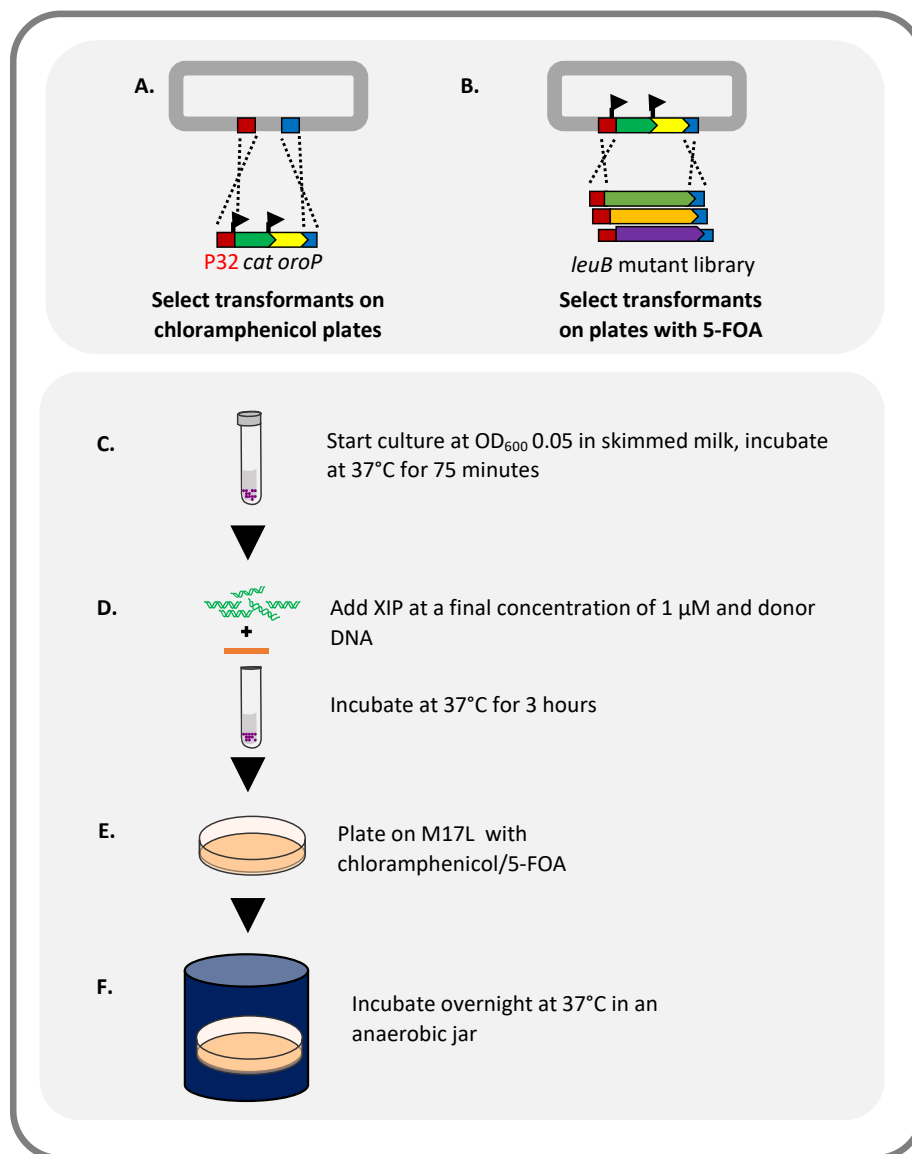


Figure 4. 2. Overview of the protocol designed for creating chromosomal libraries using natural transformation in *S. thermophilus* (A) A *cat-oroP* chimeric cassette with upstream (UP-arm) and downstream (DN-arm)

Materials and Methods Chapter 3.3

homology fragments was constructed in-vitro by Gibson assembly. The cassette was introduced in the chromosome by natural transformation and double homologous recombination. Transformants were selected for acquired resistance to chloramphenicol. **(B)** The mutant from step A was transformed with the ep-PCR cassette. Transformants were selected for loss of *oroP* gene, resulting in resistance to 5-FOA. **(C) - (F)** Schematic protocol for genetic modification using natural transformation in *S. thermophilus*.

As an example, the following protocols are described for building a library of *leuB* genes (1038 bp) created by error-prone PCR and inserted into the *tRNA^{Ser}* locus of *S. thermophilus*. To achieve this, a recipient strain containing a *cat-oroP* cassette inserted at the target locus is built. In a second step, the cassette is replaced by the library of genes using the counter selection principle.

i. Assembling the U-*cat-oroP*-D cassette

A cassette comprising *cat* and *oroP* genes under constitutive promoter control and flanked by 1.5-kb DNA fragments corresponding to the upstream (U) and downstream (D) regions of the *tRNA^{Ser}* locus is assembled (see **Note 7**). All primers used in this section can be found in **Table 4. 1** and were designed with 40 nucleotides overlapping sequences.

Building Scarless Gene Libraries in the Chromosome of Bacteria

1. Primer pairs K1 and K2 were used to amplify the U and D fragments from *S. thermophilus* genomic DNA as template, respectively (see **Note 8**).
 2. Primer pairs K3 was used to amplify the *cat* fragment from the strain described in **(25)**.
 3. Primer pair K4 was used to amplify the *oroP* fragment from plasmid pCS1966 described in (Solem et al. 2008)
 4. Clean-up the fragments by using the Monarch PCR & DNA Cleanup kit or equivalent.
 5. Mix the four purified DNA fragments to get equimolar (1 pmol) concentration, add ddH_2O to 5 μL (see **Note 9**).
 6. Incubate with 15 μL of the Gibson Master Mix (see **Note 10**) for 1 hour at 50°C.
 7. Verify the presence of the expected band of full-length cassette on agarose gel.
 8. Clean-up the product using column purification. Elute in ~10 μL of ddH_2O or elution buffer.
- ii. **Chromosomal insertion of U-*cat-oroP*-D cassette at the tRNA^{Ser} locus. Adapted from** (Fontaine, Boutry, Henry De Frahan, Delplace, Fremaux, et al. 2010)

Day 0

Materials and Methods Chapter 3.3

1. Start a preculture of *S. thermophilus* LMD9 in 1ml of CDML (see **Note 11**) at 37C without agitation.

Day 1

1. Dilute the culture in 1 ml of skimmed milk medium (SMM) to an OD₆₀₀ of 0.05 (typically 50 µL in 1 mL milk) and incubate 75 min at 37C without agitation (see **Note 12**).
2. Thaw XIP peptide (500 µM stock) and add 2 µl to the culture (1 µM final).
3. Mix 300 µL of cells in a 1.5 mL microcentrifuge tube with the purified assembled donor DNA cassette (10 µL) from 3.2.2. Take 300 µL in another microcentrifuge tube as control (without donor DNA).
4. Incubate both tubes at 37C for 3 hours without agitation.
5. Spread 100 µL of cells on M17L + chloramphenicol plates.
6. Incubate the plates O/N at 37C in an anaerobic jar.

Day 2

1. Pick several clones and re-streak them on M17L + chloramphenicol to prevent that non-resistant cells propagate in association with resistant cells (streptococci can form chains). (see **Note 13**)
2. Incubate the plate overnight at 37C in an anaerobic jar.

Day 3

Building Scarless Gene Libraries in the Chromosome of Bacteria

1. Pick an isolated colony of each clone, perform colony PCR (see section 2 for detailed protocol) to verify presence of insert and cultivate one clone to store it as a glycerol stock (Day 4).

iii. Assembling the U-*lib*-D cassette

A cassette comprising a library of *leuB* genes under its constitutive promoter and flanked by 1.5-kb DNA fragments corresponding to the upstream (U) and downstream (D) regions of the *tRNA^{Ser}* locus is assembled. All primers used in this section are in **Table 4. 1** and were designed with 40 nucleotides overlapping sequences.

1. Primer pairs K5 and K6 were used to amplify the U and D fragments, respectively.
2. Primer pair K7 was used to perform an error-prone PCR on the *leuB* gene as described in section 2. The product is called *lib*.
3. Clean up the PCR products with using the Monarch PCR & DNA Cleanup kit or equivalent.
4. Mix the three purified DNA fragments to get equimolar (1 pmol) concentration, add ddH_2O to 5 μL (see **Note 9**).
5. Add 15 μL of Gibson Assembly Master Mix and ddH_2O to 20 μL . Incubate for 1 hour at 50°C.
6. Verify the presence of the expected band of full-length cassette on agarose gel.

Materials and Methods Chapter 3.3

7. Clean-up the product using the Monarch PCR & DNA Cleanup kit or equivalent. Elute in 6 μ L of ddH_2O or elution buffer.
8. In order to obtain large variant libraries, it is necessary to prepare large amounts of linear donor DNA. To do so, the product of Gibson assembly is further amplified by a Q5-mediated PCR using primer pair K7 (see **Note 14**).

iv. Chromosomal insertion of U-*lib*-D library at the *tRNA^{Ser}* locus (Figure 4. 2)

Day 0.

1. Start a preculture of the strain *tRNA^{Ser}::cat-oroP* obtained in section ii in 1ml of CDML supplemented with chloramphenicol (5 mg/L).

Day 1

1. Dilute the culture in 1ml of milk to an OD_{600} of 0.05 and incubate 75 min at 37°C without agitation. (see **Note 12**)
2. Thaw the XIP peptide (500 μ M stock) and add 2 μ l to the culture (1 μ M final).
3. In a 1.5 mL microcentrifuge tube, mix 300 μ L of cells with ~400ng (typically higher concentrations lead to higher number of transformants) of the assembled U-*lib*-D fragment. Take 300 μ L in another microcentrifuge tube as control (without donor DNA).

Building Scarless Gene Libraries in the Chromosome of Bacteria

4. Incubate both tubes at 37°C for 3 hours without agitation.
5. Spread the cells on M17L + 5-FOA (500 mg/L) plates.
6. Incubate the plates overnight at 37°C in an anaerobic jar.

Day 2

1. Add 2 mL of CDML and scrape off the cells using a spreader.
2. Store the library with 30% glycerol at -80°C for further analysis.

4.1.5. PCR protocols

Three different protocols are used in this chapter. High fidelity polymerase reactions were performed by using Q5 polymerase, Error-Prone PCRs were performed by using the GeneMorph II Random Mutagenesis Kit, routine colony PCR for verification of inserts.

1. Q5 PCR mix: 10 µL of 5 x Q5 Reaction Buffer, 0.7 µL of 20 mM dNTPs, 2.5 µL of 10 µM Forward and reverse primers, less than 100 ng of template DNA, 1 U Q5 High-Fidelity DNA Polymerase (NEB) and nuclease free H₂O up to 50 µL.

PCR thermocycler conditions: initial denaturation at 98°C for 30 seconds; 25-30 cycles containing: denaturation at 98°C for 20 seconds; annealing at temperature calculated for each pair of primers on <https://tmcalculator.neb.com/#!/main> for 30 s; 20–30 seconds/kb extension at 72°C; a final extension for 5 minute at 72°C and hold at 4°C.

Materials and Methods Chapter 3.3

2. Error-prone PCR (ep-PCR) using GeneMorph II Random Mutagenesis Kit: 5 μ l of 10 $\times$ Mutazyme II reaction buffer, 1 μ l of 40 mM dNTP mix (provided in kit), 0.5 μ l of primer mix (250 ng/ μ l of each primer), 2.5 U of Mutazyme II DNA polymerase, 100 ng of template DNA (see **Note 15**), and nuclease free H₂O up to 50 μ L.

PCR thermocycler conditions: initial denaturation at 95C for 2 minutes; 20-25 cycles (depending on desired mutation rate, see **Note 15**): denaturation at 95C for 30 seconds; annealing 30 seconds at $T = T_m - 5C$ (T_m calculated for each pair of primers on <https://tmcalculator.neb.com/#!/main>); extension 1 minute (≤ 1 -kb targets) or 1 minute/kb (>1 -kb targets) at 72C; a final extension for 10 minutes at 72C and hold at 4C.

Optimization of mutation rate: To generate low mutation rate (between 0 and 5 mutation/kb), instructions of the GeneMorph II Random Mutagenesis kit are followed according to manufacturer. According to the kit's protocols, mutation rates can be lowered by using higher template DNA amounts and by decreasing the number of cycles of ep-PCR reaction. Therefore, for a 1 kb fragment, an initial target amount of 100 ng is required for a run with 15-20 cycles of PCR, generating a low mutation rate between 0 and 5 mutation/kb.

3. Colony PCR: 5 μ L of 5 $\times$ GoTaq Flexi Buffer, 4 μ L of 25 mM MgCl₂, 0.7 μ L of 20 mM dNTPs, 1.5 μ L of 10 μ M Forward and reverse primers, 1.25 U GoTaq DNA Polymerase (Promega) and nuclease free H₂O up to 50

Building Scarless Gene Libraries in the Chromosome of Bacteria

μL. Use a toothpick to stab a single colony and rub it inside the PCR tube to detach some cells as template.

PCR thermocycler conditions: initial denaturation at 95°C for 4 minutes (see **Note 16**); 25 cycles: denaturation at 95°C for 30 seconds; annealing at T_a (calculated for each pair of primers on <https://tmcalculator.neb.com/#!/main>) for 30 seconds; 1 minute/kb extension at 72°C; a final extension for 7 minutes at 72°C and hold at 4°C.

4.1.6. Notes

Notes 1 – 6 in the published chapter deal with library preparation in *E. coli*. Hence, these were not included in this work.

- i. Expression levels of genes inserted in different chromosomal loci vary significantly. Care should be taken when choosing an insertion site in the chromosome. Known neutral loci such as that of tRNA^{Ser} are recommended.
- ii. The length of the homology arms is important for the efficiency of transformation in *S. thermophilus*. We have found that a minimum length of 800 bp for each of the flanking homologous arm is required. Transformation efficiency improves until 1500 bp where efficiency reaches a plateau.
- iii. The NEB Bio-calculator (<https://nebiocalculator.neb.com/>) is a handy tool to calculate several parameters of DNA

Materials and Methods Chapter 3.3

oligonucleotides and fragments such as molarity, molecular weights, etc.

- iv. Alternatively, overlap extension PCR (OE-PCR) can be used to assemble DNA fragments with sufficient homology (< 20 nucleotides). The protocol for OE-PCR is already discussed in Shevchuk, 2004. The added benefit of using Gibson assembly is that multiple fragments can be assembled in a single step with high efficiency.
- v. Since *S. thermophilus* is an obligate anaerobe, it is advantageous to minimize the air volume in the culture tube. Typically, we use a minimum 1mL of media + inoculum into a 1.5 mL microcentrifuge tube. Cells are typically grown at 37°C without agitation.
- vi. All transformations were performed in milk. Typically, we have observed transformation efficiencies one order of magnitude higher in milk compared to transformations in chemically defined medium (CDM).
- vii. *S. thermophilus* grows in pairs or chains of cells, so an additional re-streaking step on the same selection medium is necessary to obtain genetically monoclonal cells after transformation.
- viii. When amplifying a library by PCR, it is important to analyze the PCR reaction every 2 or 3 cycles so that the minimal number of cycles that are necessary for obtaining a maximal PCR yield can be estimated. This can be used to estimate the number of DNA

Building Scarless Gene Libraries in the Chromosome of Bacteria

molecules that have been amplified. This number must be at least 100-fold higher than the expected final diversity of the library. Avoid any unnecessary PCR cycles that does not improve the yield since this can generate biases in the library.

- ix. The mutation rate can be controlled by adjusting the number of epPCR cycles and the amount of initial template. 100 ng template and 20-25 cycles afford a low mutation frequency (between 0 and 5 mutations/kb). See also the user manual for the “Random Mutagenesis Kit” from Agilent.
- x. Gram-positive bacteria such as *S. thermophilus* have a thick cell-wall peptidoglycan. For best results, when performing a colony PCR on *S. thermophilus* cells, make the initial denaturation step (at 98°C or 95°C) 5 minutes long.

4.2.

Materials and Methods for Chapter 3.1

Strains and growth media

As described in Section 4.1.1. Additionally CDMG without leucine (CDMG-L) and CDMG without Lysine (CDMG-K) were prepared by following the procedure described by Letort & Julliard, 2001 (Letort and Juillard 2001) and skipping the mentioned amino acid. Solid agar plates were prepared by adding 2% (wt vol⁻¹) agar to the medium

Construction of the strains

Preparation of donor DNA is described in the results. Natural transformation was performed as described in Section 4.1.4

Measurement of D-alanine, leucine auxotrophy phenotype

For measurement of D-alanine auxotrophy, 10-fold dilutions of the Δalr transformation mix were prepared in M17G immediately after transformation and 3 μ l drops were spotted onto solid M17G + Chl plates. For measurement of leucine auxotrophy, 10-fold dilutions of the $\Delta leuB$ transformation mix were prepared in CDMG-L immediately after transformation and 3 μ l drops were spotted onto solid CDMG-L+ Chl plates. Plates were incubated anaerobically for 24 hrs.

Materials and Methods for Chapter 3.1

Measurements of growth and luciferase activity

300 μ l volume of overnight culture at OD₆₀₀ of 0.05 were incubated in a white 96-well microplate with a transparent bottom (Greiner, Alphen a/d Rijn, The Netherlands). Growth (OD₆₀₀) and luciferase activity (expressed as a relative light unit; RLU) were monitored at 10 min intervals for 24 hours or less in a multi-well plate reader (Hidex Sense, Hidex, Turku, Finland) as previously described (Fontaine, Boutry, De Frahan, Delplace, Fremaux, et al. 2010).

Construction of Alr library

The *alr(strep)* gene was amplified by Q5 polymerase PCR as described in 4.1.5.1. 100 ng of Q5 amplified *alr(strep)* gene was used as template for error-prone PCR (ep-PCR) using GeneMorph II Random Mutagenesis Kit as described in 4.1.5.2. 22 PCR cycles were performed to optimize the mutation rate between 0 and 4.5 mutations/kb. The ep-PCR product was assembled with homologous arms corresponding to the *alr* locus by Gibson assembly. This was used as donor DNA for natural transformation in 10 ml of SMM as described in Section 4.1.4. 1 ml of the transformation mix was spread each on 10 large petri plates (Corning® 500cm² square culture dish) with M17G, 1.5mg/ml 5-FOA, 0.5 mg/ml D-alanine supplementation. 10-fold dilutions of the transformation mix were prepared and 3 μ l droplets were spotted onto M17G plates with 1.5mg/ml 5-FOA, 0.5 mg/ml D-alanine supplementation to estimate the number of

Materials and Methods Chapter 3.3

transformants. Large 500 cm² plates were incubated anaerobically by sealing them in air-tight polyethylene bags with BBL GasPak systems (Becton Dickinson, Franklin lakes, NJ) at 37°C for 24 hrs. 5 ml M17G was added to the plate and grown cells were scraped and recovered. 1 ml was aliquoted in TPP[®] cryotubes (Sigma Aldrich) and stored at -80°C. 10-fold dilutions were prepared and 3 µl were spotted onto M17G plates with 0.5 mg/ml D-alanine supplementation to estimate the concentration of cells in the cryo-stock.

Deletion of *lysA* from a library of Alr variants

Roughly 3.29×10^9 of cells from the Alr library cryo-stock was inoculated into 200 ml of SMM + D-alanine. Cells were allowed to recover for 1 hour. 40 µg of $\Delta lysA::cat$ donor DNA was used for natural transformation in as described in Section 4.1.4. Cells after transformation were harvested by centrifugation (10 min, 5000 x g) and resuspended in 10 ml SMM and 1 ml of the resuspended cells were spread each on 10 large M17G + 5 µg/ml Chl + 0.5 mg/ml D-alanine petri plates (Corning[®] 500cm² square culture dish). 10-fold dilutions of the transformation mix were prepared and 3 µl droplets were spotted onto M17G plates with 5 µg/ml Chl, 0.5 mg/ml D-alanine supplementation to estimate the number of transformants. Cells were grown and recovered from the large plates as described above. 10-fold dilutions were prepared and 3 µl droplets were spotted onto CDMG-

Materials and Methods for Chapter 3.2

K plates with 5 µg/ml Chl, 0.5 mg/ml D-alanine supplementation to estimate the concentration of lysine auxotrophs.

4.3.

Materials and Methods for Chapter 3.2

Strains and growth media

As described in Section 4.1.1.

Construction of Alr library

Described in Section 4.2

Measurements of growth and luciferase activity

As described in Section 4.2

Experimental evolution of Alr libraries in condition C_1 and C_2

Cells corresponding to 15-times the size of the library ($15 * 3.29 \times 10^8 = 4.93 \times 10^9$) were inoculated into 1000 ml of M17G (C_1) and M17G + 0.5 mg/ml D-alanine (C_2) in 1000 ml bottles (SCHOTT DURAN®, Germany). Care must be taken to minimise air volume in bottles for optimum anaerobic growth conditions. Cells were stored at 37 °C. At each 24 hrs interval, 10 ml of cells were passaged into fresh media and 10 ml of cells were collected stored at -80 °C for further analysis. 10-fold dilutions were

Materials and Methods Chapter 3.3

prepared and 3 μ l droplets plated each day to ensure that a minimum of 3.29×10^9 cells were transferred to avoid bottlenecking the library diversity.

Phenotypic characterization of experimentally evolved Alr library

Due to practical considerations (too laborious for one day), it was not possible to simultaneously characterize the change in the percentage of active Alr in the library at each day of the experimental evolution. DNA from the cells collected at each day of the experimental evolution was re-transformed into the parent strain ($\Delta alr::P_{32-cat-P_{sp-oroP}}$). Dilutions of the transformed cells were plated onto M17G plates (active colonies) and M17G +D-alanine (total colonies), to calculate the Alr percent active using equation (1) (Chapter 3.2). The graph was prepared using Prism GraphPad 9.

Sanger sequencing of individual colonies

Single colonies were picked and re-streaked to break any clumps (as described in Section 4.1) and amplified by Q5 PCR. Thermocycler conditions were, initial denaturation at 98°C for 5 minutes; 25-30 cycles containing: denaturation at 98°C for 20 seconds; annealing at temperature calculated for each pair of primers on <https://tmcalculator.neb.com/> for 30 s; 45 seconds/kb extension at 72°C; a final extension for 5 minute at 72°C and hold at 4°C.

Materials and Methods for Chapter 3.2

The longer initial denaturation is necessary for the Gram positive *S. thermophilus* cells to be lysed.

NGS sequencing

Genomic DNA extraction

Genomic extraction was performed by using the Monarch Genomic DNA Purification Kit - Genomic DNA Purification from Gram-positive Bacteria and Archaea, NEB by following the manufacturer's recommendations. 5 units of mutanolysin enzyme (Sigma-Aldrich) were additionally added for improved lysis of *S. thermophilus* cells.

Preparation of library DNA for Ion torrent sequencing

For Ion torrent sequencing, 10 ml of cells from each time point in the library was pelleted, lysed and the genome was extracted as described in Section 6.3.1 above. 2 µg of genomic DNA (corresponding to 1.050×10^9 copies of *S. thermophilus* 1.8 Mb genome) was used as input for PCR with the Q5 polymerase as described in section 6.1.3 above. The PCR product was electrophoresed on a 0.8% agarose gel and subsequently excised and purified using the Monarch DNA Gel Extraction Kit from New England Biolabs (NEB). This was subsequently quantified using NanoDrop and Qubit dsDNA HS Assay. The pure DNA was submitted to the lab of Laboratory of Human Molecular Genetics, de Duve Institute, UCLouvain, which processed them using the Ion Xpress Plus Fragment Library kit from

Materials and Methods Chapter 3.3

ThermoFisher Scientific according to manufacturer's instructions and sequenced them on a Ion Torrent PGM dividing the library samples between two parallelly run chips.

Ion torrent sequencing data analysis

In this section, we have the details of the processing steps of the Ion torrent sequencing data analysis pipeline. An in-house written script in python was used to perform the analysis (script - Appendix A).

The data generated by the Ion Torrent PGM was received as unmapped FASTQ files, with the sequence information and per base quality scores (Q value). We did not find any software package to read and pre-process Ion-torrent reads for variant calling, most of the existing software packages like Enrich (Fowler et al. 2011) were optimised for Illumina generated sequencing reads. So we wrote in-house python scripts to analyse the reads for variant calling (the unmodified scripts are shared in the appendix). The pipeline of the analysis is as follows.

Trimming and Filtering: First, the FASTQ file was filtered to remove reads shorter than 65 nucleotides. Next, reads where the average per base sequence quality (Q) was less than a value of 25 were discarded. Next, the reads were trimmed to remove reads on the extremities with a Q value less than 25.

Materials and Methods for Chapter 3.2

Alignment and Mapping to template: The trimmed and filtered reads were locally aligned to the template 1156 bp alanine racemase sequence using the Smith-Waterman algorithm with the parameters given below. Poorly aligned reads with a score below or equal to 30 were discarded.

Algorithm: [Smith-Waterman Pairwise](#) sequence aligner

```
match_score: 1.000000
mismatch_score: -2.000000
target_open_gap_score: -2.500000
target_extend_gap_score: -2.500000
target_left_open_gap_score: -2.500000
target_left_extend_gap_score: -2.500000
target_right_open_gap_score: -2.500000
target_right_extend_gap_score: -2.500000
query_open_gap_score: -2.500000
query_extend_gap_score: -2.500000
query_left_open_gap_score: -2.500000
query_left_extend_gap_score: -2.500000
query_right_open_gap_score: -2.500000
query_right_extend_gap_score: -2.500000
mode: local
```

Translation to protein: The reading frame of the reading frame for the aligned reads were calculated and trimmed appropriately. Subsequently, the translated reads were mapped to the translated template sequence.

Frequency and enrichment calculation (script - Appendix B)

The next step involved the computation of the depth of sequencing at each amino acid position, and count of the number of times a substitution was observed. Frequency was calculated as the ratio between the number of counts observed for each substitution and the depth of sequencing. This is given in equation I. Enrichment ratios were calculated by normalizing the frequency for the substitution by the counts observed for

Materials and Methods Chapter 3.3

the wild-type amino acid at that position. This is given in equation II. Substitutions for which counts were observed 10 times or less were considered to be statistically irrelevant and masked, these substitutions were not included in any subsequent calculations.

Graphical representations of the Alr experimental evolution E_m^r values

The python package dmslogo (<https://jbloomlab.github.io/dmslogo/>) was modified (Script - Appendix D) to create the representation of modified E_m^r values as a logo plot (**Figure 2. 10** and **Figure 2. 22**). Representations of fitness distributions were prepared by sorting the E_m^r values into bins of 0.05. The visualizations (**Figure 2. 11**, **Figure 2. 12**, **Figure 2. 15**, **Figure 2. 14**, **Figure 2. 23**, **Figure 2. 24**, **Figure 2. 25**, **Figure 2. 26**) were generated using Prism GraphPad 9. Scores for predicted impact of substitutions were obtained by using the PROVEAN and SIFT algorithm. The sequence of wild-type AlrSt was provided as input and scores were calculated using default parameters. PROVEAN is available on the homepage of the J. Craig Venter Institute at <http://provean.jcvi.org>. SIFT is available at <https://sift.bii.a-star.edu.sg/>.

4.4.

Materials and Methods for Chapter 3.3

Strains and growth media

As described in Section 4.1.1.

Materials and Methods for Chapter 3.3

Constructing AlrSt insertants with SARS-CoV-2 protease recognition sites

alr gene was replaced by a P₃₂-cat P_{SP}-oroP cassette, transformants were selected on M17G supplemented with 5 µg/ml Chl + D-alanine. Subsequently, P₃₂-cat P_{SP}-oroP was replaced by the *alr*^{insert} library, transformants were selected on M17G with 5-FOA (2mg/ml).

Construction of the *plpro* Δ *alr::alr*^{*plpro*} and *mpro* Δ *alr::alr*^{*mpro*} strains

A cassette consisting of P₃₂-cat P_{SP}-oroP P_{Xyl2} pGyr-*xylR* was introduced at the tRNA_{serine} locus by natural transformation, transformants were selected on M17G supplemented with 5 µg/ml Chl. *mpro* (921bp) and *plpro* (5838 bp) were amplified individually from a plasmid collection (Addgene) encoding the different coding regions of SARS-CoV-2 (D. K. Kim et al. 2020) and cloned downstream the P_{Xyl2} promoter (Xie, Qi, and Merritt 2013) by replacement of the P₃₂-cat P_{SP}-oroP genes by natural transformation and selecting transformants on M17G with 2 mg/ml 5-FOA and 0.5 mg/ml D-alanine. The protease sequences were designed to include the N-terminal self-cleavage site required for the protease maturation (L. Zhang et al. 2020).

Construction of the SICLOPS libraries

A cassette consisting *ΔleuB::P₃₂-cat P_{SP}-oroP* was introduced into the *plpro* Δ *alr::alr*^{*plpro*} and *mpro* Δ *alr::alr*^{*mpro*} strains by natural transformation and selection on M17G supplemented with 5 µg/ml Chl + 0.5 mg/ml D-

Materials and Methods Chapter 3.3

alanine. Four SICLOPPS libraries were designed using oligonucleotide primers (Eurogentec, Belgium) with degenerated codons. We used a similar strategy for the construction of the first three libraries, a cysteine followed by three, five or seven randomised codons, creating the libraries with the motif CX₃, CX₅ and CX₇ (X = NNK, where N is A, T, C, or G; and K is G or T). For the fourth library, we included the PLpro recognition site (LXGG) in the design through a semi degenerated codon library, creating the motif CX₃LZGGX₂ (Z = ARV, where R is G or A; and V is G, C or A). ARV degeneracy was used for the 6th position because Arginine, Lysine and Asparagine are the commonly observed amino acids in SARS-CoV-2 PLpro cleavage sites. The partial degeneracy scheme NNK was used instead of full NNN degeneracy to reduce the number of stop codons encoded and allow a more uniform distribution of amino acids (Cheng et al. 2007). Primers containing the degenerated codons were used to amplify the regions flanking the *tRNA_{serine}* locus, the PCR products were assembled using Gibson assembly and this was used as donor DNA for natural transformation to replace the *cat-oroP* cassette at the *leuB* locus. Transformants were selected on 1.5 mg/ml 5-FOA and 0.5 mg/ml D-alanine. Cells were collected and stored as described for the Alr library in Section 4.2.

Screening the SICLOPS libraries for PLpro inhibitory peptides

Materials and Methods for Chapter 3.3

The four strains *plpro* $\Delta alr::alr^{plpro}$ $\Delta leuB :: SICLOPS^{lib}$ were screened for PLpro inhibitory peptides by plating 10^9 cells of each library on M17G + 1.5 mg/ml 5-FOA + 0.25% xylose. To reduce frequency of false positives, transformant cells were collected and lysed, the genome was extracted (as described in Section 4.3), and the *leuB* locus encoding *SICLOPS^{lib}* and 1.5kB flanking homologous arms were amplified by Q5 PCR (as described in Section 4.1) The amplified product was used as donor DNA for natural transformation (as described in Section 4.1) of the parent strain (*plpro* $\Delta alr::alr^{plpro}$ $\Delta leuB::P_{32-cat}$ $P_{SP-oroP}$ strain), where transformants were selected on M17G + 1.5 mg/ml 5-FOA + 0.25% xylose. This cycle was repeated four times and cells collected were prepared for NGS sequencing (as described in Section 4.2).

Materials and methods: Appendices

Here shared are the python codes which were used to process the raw Ion-torrent sequencing data. Subsequently code used to generate images are shared.

A. Filtering, trimming and alignment

```

1. import os
2. import numpy as np
3. import glob
4. from Bio import SeqIO
5.
6. from Bio.Seq import Seq
7. from Bio import Align
8.
9. # from Bio.Alphabet import IUPAC
10.
11. files = glob.glob('*.fastq')
12.
13. for f in files:
14.     folder = f[:-6]
15.     os.makedirs(folder)
16.     aa = open(f'{folder}/aa.txt', 'wt') #ouvre le fichier de sortie d'alignement des aa
17.     parameters_align1 = open(f'{folder}/parameters_align1.txt', 'wt')
18.
19. #-----
20. #ALIGNMENT
21.
22.     aligner = Align.PairwiseAligner()
23.     aligner.mode = 'local'
24.     aligner.match_score = 1.0
25.     aligner.mismatch_score = -2.0
26.     aligner.gap_score = -2.5
27.
28.     wt=Seq('atgattgaggtttacatcgaccaaccttggcaaaggttgatctgtcagcaatttcagaaaattgaacaagtggt
tagtcatattcctaaacaggtgcagacttttcagttgtaaggccaatgcttatggtcatggcgctgttgaaagtggtctaaacatgtgt
ctaagcaggtggatgggttttgtgttctaacctagatgaggcgcttgaaacttcgtcaagctggcatagagcaaccgatacttatttg
ggagtagtcttccagatgggtgtcccttggtctattcaagaaaacattagctgacgggtggcctcacttgagtggtggcactggctca
gaacaagggttgatttgataggttgacgtgcatattaaggttgattcaggtatgggtcgtattgggtacgtaacctaaaggac
gcagataacttgattgctggtctgaaggctcttggtgctgatgtcgaaggatttttactcattttgcaacagctgatgaagcagatga
tagtaaaattgaagcgtcaattatcttctcaccgatttggtggataacttgactgataagctcgcttggtgcacgctagcaactcagc
gacaagtatttgcatgcagcaaccgtcttaatacggtagctctgggtgttgatctatggtctcaatcctagtggcagtggtcttgga
acttccctataatattcaacctgccctaagtttagagacagctcttattcatgttaagacccttcagccggtcaagacgtgggttacg
gtgcaacctatacaactacggctgaagaagttattggaacgttcgattggctatgctggtggtgactagagatttacaaggtt
ccatgttatttggtggatggacaactttgtctattgttggcggtgtctatggaccagatcactgttcgcttcctaaggtttatccacta

```

Materials and methods: Appendices

```

ggaactccagttacgcttatgggtgaaatggaggagcgagcattactgcgacagagggtggctgaaaagcgtggcacaatcaact
acgaggtcctttgttacttagtgatcggtacctcgtagctatgacgggtcctcgggtagcggaggctcgggttgagccaccgcag
ttcgaataataa')
29.     seq1 = wt #<- choisi le template
30.
31.     def Alignment(seq1, seq2):
32.         score = aligner.score(seq1, seq2)
33.         score_rev = aligner.score(seq1, seq2.reverse_complement())
34.         if score > score_rev:
35.             alignments = aligner.align(seq1, seq2)
36.
37.         else:
38.             alignments = aligner.align(seq1, seq2.reverse_complement())
39.             score = score_rev
40.
41.         return score, alignments[0]
42.
43. #-----
44.
45.
46. #Création de la liste des fichiers FASTA présnt dans le répertoire courant par os.listdir('.')
47.     nbre_raw = 0
48.     short_seq = 0
49.     bad_quality = 0
50.     size_shortseq = 65
51.     threshold_meanQuality = 25
52.     mean_quality = 0
53.     toshort = 0
54.     mismatching = 0
55.     rest = 0
56.     shiftL_edition = [0,2,1] # Pour conversion la division euclidienne lors du calcul de shift a
    gauche avant la traduction de la séquence
57.     splitfile_dna = 0
58.
59.     for seq_record in SeqIO.parse(f, "fastq"):
60.         left = 0
61.         right = len(seq_record)
62.         nbre_raw += 1
63.         #print(nbre_raw)
64.         if len(seq_record) < size_shortseq:
65.             short_seq += 1
66.         elif np.mean(seq_record.letter_annotations["phred_quality"]) <
threshold_meanQuality:
67.             mean_quality += 1
68.         else:
69.             for n in range(len(seq_record)):
70.                 if
seq_record.letter_annotations["phred_quality"][n] < threshold_meanQuality:
71.                     left = n
72.                 else:

```

Materials and Methods Chapter 3.3

```

73.                                     break
74.         for n in range(1,len(seq_record)-1):
75.             if
seq_record.letter_annotations["phred_quality"][-n] < threshold_meanQuality:
76.                 right = len(seq_record)-n
77.             else:
78.                 break
79.         edited = seq_record[left:right]
80.
81.         if len(edited.seq) > 30:
82.             score, alignment = Alignment(seq1, edited.seq)
83.             if score >= 30:
84.                 rest += 1
85.                 print(seq_record.id, nbre_raw)
86.                 print('score = ' + str(score))
87.                 shiftL = alignment.aligned[1][0][0] +
shiftL_edition[alignment.aligned[0][0][0]%3] # démarre a la première base qui est allinée +
le reste de la division euclidienne histoire de démarrer sur un codon
88.                 shiftR = alignment.aligned[1][0][-1] -
(alignment.aligned[1][0][-1] - shiftL)%3 # retire 0, 1 ou 2 bases pour arriver à un multiple de 3
89.                 alignment_position =
(alignment.aligned[0][0][0] + shiftL_edition[alignment.aligned[0][0][0]%3])/3 # position du
premier aa
90.                 seqDNA =
Seq(alignment.query[shiftL:shiftR])#, IUPAC.unambiguous_dna)
91.                 seqPROT = seqDNA.translate()
92.                 seqPROTnormalise = str('-
'*alignment_position+str(seqPROT)+'-'*(len(seq1.translate())-len(seqPROT)-
alignment_position))
93.                 print(seqPROTnormalise, file=aa)
#ecrit dans le fichier d'alignement des séquences nucléiques
94.
95.
96.
97.                 for n in range(len(edited)):
98.                     if
edited.letter_annotations["phred_quality"][n] <= 13:
99.                         bad_quality += 1
100.                                     break
101.
102.                                     #del edited
103.                                     #del shiftL
104.                                     #del shiftR
105.                                     #del alignment_position
106.                                     #del seqDNA
107.                                     #del seqPROT
108.                                     #del seqPROTnormalise
109.                                     #del score
110.                                     #del alignment

```

Materials and methods: Appendices

```
111.
112.                                     else:
113.                                     mismatching += 1
114.                                     else:
115.                                     toshort += 1
116.                                     #del seq_record
117.
118.
119.
120.
121.
122.     print(f'Number of raw data = {nbre_raw}', file=parameters_align1)
123.     print(f'Remove {short_seq} sequences < {size_shortseq} nt', file=parameters_align1)
124.     print(f'Remove {mean_quality} sequences which have < {threshold_meanQuality} of
mean quality', file=parameters_align1)
125.     print(f'Too short for alignment = {toshort}', file=parameters_align1)
126.     print(f'Mismatching = {mismatching}', file=parameters_align1)
127.     print(f'Rest {rest} sequence(s)', file=parameters_align1)
128.     print(f'{bad_quality} sequence(s) with 1 or more nucleotide(s) with poor quality',
file=parameters_align1)
129.     print(f'Algorithm: {aligner.algorithm}', file=parameters_align1)
130.     print(aligner, file=parameters_align1)
131.
132.     aa.close()
133.     parameters_align1.close()
```

B. Frequency calculation

```
1. import csv
2. import numpy as np
3. from Bio.Seq import Seq
4. import seaborn as sns
5. import matplotlib.pyplot as plt
6.
7. alignment = open("aa.txt", "r")
8.
9. seq1 =
Seq('atgattgcgagtttacatcgaccaacctggcaaggtgatctgtcagcaatttcagaaaaattgaacaagtggttagtcat
attcctaaacaggtgcagacttttcagttgtaaggccaatgcttatggctatggcgctgttgaaagtggttaaacatgtgtctaagc
aggtggatggtttttgtgtttctaacctagatgaggcgcttgaaacttcgtcaagctggcatagagcaaccgatacttattttggagta
gtcttgcagatggtgttcccttggtctattcaagaaaacattagctgcggtggcctcacttgagtgttggtcactggctcagaaaca
agggcttgattgataggttgacctgtcatattaaggttgattcaggtatgggtcgtattgggtacgtaacctaaaggacgcagat
aacttgattgctggtctgaaggctcttggtgctgatgtcgaaggtattttactcattttgcaacagctgatgaagcagatgatagtaa
atttaagcgtcaattatcttttcaccgatttggtgataacttgactgatagacctgcttggtgcacgctagcaactcagcgacaa
gtatttgcatgcagcaaccgtcttaatacgggtacgtctgggtgtgtcatctatggtctcaatcctagtggtgagttgttgaaactc
cctataatattcaacctgccctaagtttagagacagctcttattcatgttaagaccttcacgggtcaagacgtgggttacgggtgca
acctatacaactacggctgaagaagtatttggaacgttcgattggctatgcggatggttgactagagatttacaaggtttccatg
ttattggtgaggaacattgtcctattgttggcggtgtgtctatggaccagatcactgttcgcttcctaaggtttaccactaggaac
```

Materials and Methods Chapter 3.3

```
    tccagttacgcttatgggtgaaaatggaggagcgagcattactgcgacagaggtggctgaaaagcgtggcacaatcaactacgag
    gtcctttgtttacttagtgatcgtgtacctcgtagctatgacggttcctcgggtagcggaggctcgggttggagccaccgcagttcga
    aaaataa')
10.
11. aa1 = seq1.translate()
12. position = list(range(1, len(aa1)+1))
13. progress = 0
14.
15. total_read = [0]*len(aa1)
16. asp = [0]*len(aa1)
17. thr = [0]*len(aa1)
18. ser = [0]*len(aa1)
19. glu = [0]*len(aa1)
20. pro = [0]*len(aa1)
21. gly = [0]*len(aa1)
22. ala = [0]*len(aa1)
23. cys = [0]*len(aa1)
24. val = [0]*len(aa1)
25. met = [0]*len(aa1)
26. ile = [0]*len(aa1)
27. leu = [0]*len(aa1)
28. tyr = [0]*len(aa1)
29. phe = [0]*len(aa1)
30. his = [0]*len(aa1)
31. lys = [0]*len(aa1)
32. arg = [0]*len(aa1)
33. trp = [0]*len(aa1)
34. gln = [0]*len(aa1)
35. asn = [0]*len(aa1)
36. stop = [0]*len(aa1)
37.
38. for line in alignment:
39.     progress += 1
40.     print('Sequences:', progress, end="\r")
41.     for i in range(len(aa1)):
42.         if line[i] != '-':
43.             total_read[i] = int(total_read[i]) + 1
44.         if line[i] != '-' and line[i] != aa1[i] and line[i] == 'D':
45.             asp[i] = int(asp[i]) + 1
46.         if line[i] != '-' and line[i] != aa1[i] and line[i] == 'T':
47.             thr[i] = int(thr[i]) + 1
48.         if line[i] != '-' and line[i] != aa1[i] and line[i] == 'S':
49.             ser[i] = int(ser[i]) + 1
50.         if line[i] != '-' and line[i] != aa1[i] and line[i] == 'E':
51.             glu[i] = int(glu[i]) + 1
52.         if line[i] != '-' and line[i] != aa1[i] and line[i] == 'P':
53.             pro[i] = int(pro[i]) + 1
54.         if line[i] != '-' and line[i] != aa1[i] and line[i] == 'G':
55.             gly[i] = int(gly[i]) + 1
56.         if line[i] != '-' and line[i] != aa1[i] and line[i] == 'A':
```

Materials and methods: Appendices

```
57.         ala[i] = int(ala[i]) + 1
58.     if line[i] != '-' and line[i] != aa1[i] and line[i] == 'C':
59.         cys[i] = int(cys[i]) + 1
60.     if line[i] != '-' and line[i] != aa1[i] and line[i] == 'V':
61.         val[i] = int(val[i]) + 1
62.     if line[i] != '-' and line[i] != aa1[i] and line[i] == 'M':
63.         met[i] = int(met[i]) + 1
64.     if line[i] != '-' and line[i] != aa1[i] and line[i] == 'I':
65.         ile[i] = int(ile[i]) + 1
66.     if line[i] != '-' and line[i] != aa1[i] and line[i] == 'L':
67.         leu[i] = int(leu[i]) + 1
68.     if line[i] != '-' and line[i] != aa1[i] and line[i] == 'Y':
69.         tyr[i] = int(tyr[i]) + 1
70.     if line[i] != '-' and line[i] != aa1[i] and line[i] == 'F':
71.         phe[i] = int(phe[i]) + 1
72.     if line[i] != '-' and line[i] != aa1[i] and line[i] == 'H':
73.         his[i] = int(his[i]) + 1
74.     if line[i] != '-' and line[i] != aa1[i] and line[i] == 'K':
75.         lys[i] = int(lys[i]) + 1
76.     if line[i] != '-' and line[i] != aa1[i] and line[i] == 'R':
77.         arg[i] = int(arg[i]) + 1
78.     if line[i] != '-' and line[i] != aa1[i] and line[i] == 'W':
79.         trp[i] = int(trp[i]) + 1
80.     if line[i] != '-' and line[i] != aa1[i] and line[i] == 'Q':
81.         gln[i] = int(gln[i]) + 1
82.     if line[i] != '-' and line[i] != aa1[i] and line[i] == 'N':
83.         asn[i] = int(asn[i]) + 1
84.     if line[i] != '-' and line[i] != aa1[i] and line[i] == '*':
85.         stop[i] = int(stop[i]) + 1
86.
87. #####
88. # Output mutation
89. #####
90.
91. with open('mutation.csv', mode='w') as mutation_file: # ouvre le .csv
92.     mutation_file = csv.writer(mutation_file, delimiter=',', quotechar='"',
93.     quoting=csv.QUOTE_MINIMAL) # format des séparateurs et paramètres du .csv
94.     mutation_file.writerow(['aa position', 'Total number of reads', 'D', 'T', 'S', 'E', 'P', 'G', 'A',
95.     'C', 'V', 'M', 'I', 'L', 'Y', 'F', 'H', 'K', 'R', 'W', 'Q', 'N', 'stop']) # écrit les entêtes
96.     mutation_file.writerows(tuple(zip(position, total_read, asp, thr, ser, glu, pro, gly, ala, cys,
97.     val, met, ile, leu, tyr, phe, his, lys, arg, trp, gln, asn, stop)))
98.
99. #####
100. # Fréquence normalisée et treeshold
101. #####
102.
103. # renvoi une frequence si le nbre de reads à cette position est suffisant
104. # ainsi que le nombre de hit pour chaque substitution, sinon renvoi la valeur 1000
105.
106. threshold_read = 5000
```

Materials and Methods Chapter 3.3

```
104. threshold_substitution = 10
105. freq = [asp, thr, ser, glu, pro, gly, ala, cys, val, met, ile, leu, tyr, phe, his, lys, arg, trp, gln, asn,
stop]
106. for n in range(len(aa1)):
107.     if total_read[n] >= threshold_read:
108.         for residu in freq:
109.             if residu[n] >= threshold_substitution:
110.                 residu[n] = int(residu[n])/int(total_read[n])
111.             else:
112.                 residu[n] = 1000
113.
114.     else:
115.         for residu in freq:
116.             residu[n] = 1000
117. #####
118. # Output frequency
119. #####
120.
121. with open('frequency.csv', mode='w') as frequency_file: # ouvre le .csv
122.     frequency_file = csv.writer(frequency_file, delimiter=',', quotechar='"',
quoting=csv.QUOTE_MINIMAL) # format des séparateurs et paramètres du .csv
123.     frequency_file.writerow(['aa position', 'Total number of reads', 'D', 'T', 'S', 'E', 'P', 'G', 'A',
'C', 'V', 'M', 'I', 'L', 'Y', 'F', 'H', 'K', 'R', 'W', 'Q', 'N', 'stop']) # écrit les entêtes
124.     frequency_file.writerows(tuple(zip(position, total_read, asp, thr, ser, glu, pro, gly, ala, cys,
val, met, ile, leu, tyr, phe, his, lys, arg, trp, gln, asn, stop)))
125.
126. #####
```

C. Calculation of Pearson correlation values between the replicates

```
1. import csv
2. import numpy as np
3. from numpy import NaN
4. import seaborn as sns
5. import matplotlib.pyplot as plt
6. from Bio.Seq import Seq
7. import pandas as pd
8.
9. seq1 =
Seq('atgattgaggtttacatcgaccaacctggcaaaggtgatctgtcagcaatttcagaaaattgaacaagtggttagtcat
attcctaaacagggtgcagacttttcagttgtaaaggccaatgcttatggatggcgctgttgaagtggtacaaacatgtgtcctaagc
aggtggatggtttttgtgttttaacctagatgaggcgcttgaacttcgtcaagctggcatagagcaaccgatacttattttgggagta
gtcttgccagatggtgtcccttggtctattcaagaaaacattagtctgacgggtggcctcacttgagtgttggtggtcactggctcagaaaca
agggtggtgattgatagggttgacctgtcatattaaggttgattcaggtatgggtcgtattgggtacgtaacctaaaggacgcagat
aacttgattgctggtcgaaggctcttggtgctgatgtcgaaggtatttttactcattttgcaacagctgatgaagcagatgatagtaa
attaagcgtcaattatcttttaccgattggtggataactgactgatagacctgcttggtgcacgctagcaactcagcgacaa
gtatttggcatgcagcaaccgtcttaatacggtagctgtgggtgtgtcatctatggtctcaatcctagtggcagtgcttggaacttc
```

Materials and methods: Appendices

```

cctataatattcaacctgcctaagtttagagacagctcttattcatgttaagacccttcagccggtcaagacgtgggttacggtgca
acctatacaactacggctgaagaagtatttggaacgcttcgattggctatgcggatgggtggactagagattacaaggtttccatg
ttattgtggatggacaactttgtcctattgttggcggtgtgtctatggaccagatcactgttcctcctaaggtttatccactaggaac
tccagttacgcttatgggtgaaaatggaggagcgagcattactgcgacagaggtggctgaaaagcgtggcacaatcaactacgag
gtcctttgttacttagtgatcgtgtacctcgtagctatgacggttcctcgggtagcggaggctcgggttgagccacccgcagttcga
aaaataa')
10. appel = ["asp", "thr", "ser", "glu", "pro", "gly", "ala", "cys", "val", "met", "ile", "leu", "tyr",
            "phe", "his", "lys", "arg", "trp", "gln", "asn", "stop"]
11. # mut = int(input('Type the position of amino acid: '))
12.
13. # import data from CSV
14.
15. def import_freq(f, col):
16.     data = np.genfromtxt(fname='frequency_0'+str(f)+'.csv', delimiter=',', dtype=float,
                            usecols=(col), skip_header=1, encoding='UTF-8')
17.     return data
18.
19. def Frequency(f):
20.     #position = np.genfromtxt(fname='frequency_'+str(f)+'.csv', delimiter=',', dtype=int,
                                usecols=(0), skip_header=1, encoding='UTF-8')
21.     #total_read_after = import_freq(f, 1)
22.     asp = import_freq(f, 2)
23.     thr = import_freq(f, 3)
24.     ser = import_freq(f, 4)
25.     glu = import_freq(f, 5)
26.     pro = import_freq(f, 6)
27.     gly = import_freq(f, 7)
28.     ala = import_freq(f, 8)
29.     cys = import_freq(f, 9)
30.     val = import_freq(f, 10)
31.     met = import_freq(f, 11)
32.     ile = import_freq(f, 12)
33.     leu = import_freq(f, 13)
34.     tyr = import_freq(f, 14)
35.     phe = import_freq(f, 15)
36.     his = import_freq(f, 16)
37.     lys = import_freq(f, 17)
38.     arg = import_freq(f, 18)
39.     trp = import_freq(f, 19)
40.     gln = import_freq(f, 20)
41.     asn = import_freq(f, 21)
42.     stop = import_freq(f, 22)
43.
44.     freq_mutation = {"asp":asp, "thr":thr, "ser":ser, "glu":glu, "pro":pro, "gly":gly, "ala":ala,
                       "cys":cys, "val":val, "met":met, "ile":ile, "leu":leu, "tyr":tyr, "phe":phe, "his":his, "lys":lys,
                       "arg":arg, "trp":trp, "gln":gln, "asn":asn, "stop":stop}
45.     return freq_mutation
46.
47. # importation des frequences
48. sample_0=Frequency(49)

```


Materials and Methods Chapter 3.3

```
49. sample_1=Frequency(50)
50. sample_2=Frequency(51)
51. sample_3=Frequency(52)
52. sample_4=Frequency(53)
53. sample_5=Frequency(54)
54. sample_6=Frequency(55)
55. sample_27=Frequency(56)
56. sample_28=Frequency(57)
57. sample_29=Frequency(58)
58. sample_30=Frequency(59)
59.
60.
61. # Normalize
62.
63. position = np.genfromtxt(fname='frequency__0'+str(49)+'.csv', delimiter=',', dtype=int,
64.                             usecols=(0), skip_header=1, encoding='UTF-8')
65. def Normalization(t0, drift):
66.     normalization = {}
67.     na = {}
68.     for i in t0:
69.         normalization[i]=[]
70.         #na[i]=[]
71.         for k in range(len(position)):
72.             if t0[i][k] != 1000 and drift[i][k] != 1000:
73.                 #if t0[i][k] != 1000 and basal[i][k] !=1000 and ind[i][k] != 1000:
74.
75.                 #normalization[i].append((ind[i][k]/t0[i][k])/(basal[i][k]/t0[i][k]))
76.                 normalization[i].append(drift[i][k]/t0[i][k])
77.                 #na[i].append(" ")
78.             else:
79.                 normalization[i].append(NaN)
80.                 #na[i].append("")
81.
82.     return normalization
83.
84. #####
85. xlabel = ['- DA rep 1', '- DA rep 1', '+ DA rep 1', '+ DA rep 1']
86. ylabel = ['+ DA rep 2', '- DA rep 2', '+ DA rep 2', '- DA rep 2']
87. sampleA = [sample_27, sample_27, sample_29, sample_29]
88. sampleB = [sample_30, sample_28, sample_30, sample_28]
89.
90.
91. #compute Pearson correlation
92.
93. pearsoncorrel = list()
94.
95. for i in range(4):
96.
```

Materials and methods: Appendices

```
97. # this is a dictionary with the amino acid as key and normalization values in it
98.     fitA = Normalization(sample_0, sampleA[i])
99.     fitB = Normalization(sample_0, sampleB[i])
100.
101. # here we pass the values in the dictionary to a dictionary of values
102.     valuesdictA = fitA.values()
103.     valuesdictB = fitB.values()
104.
105. # create new lists
106.     valuesA = list()
107.     valuesB = list()
108.
109. # pass the values in the dictionary of values to a list
110.     for x in valuesdictA:
111.         valuesA += x
112.     for x in valuesdictB:
113.         valuesB += x
114.
115. # create a pandas dataframe with all the values in rows
116.     df1 = pd.DataFrame([valuesA]).astype(float)
117.     df2 = pd.DataFrame([valuesB]).astype(float)
118.
119. # we create a list to add the correlation values
120.     pearsoncorrelation = list()
121.
122. # calculate the correlation values, axis=1 is for rows, drop= false is for ignoring nan values
123. # this creates a pandas series object
124.     pearsoncorrelation = df1.corrwith(df2,axis=1, drop=False, method ='pearson')
125.
126. # create a new list from the pandas series object
127.     pearsoncorrelation = pearsoncorrelation.tolist()
128.
129. # round the values to 2 decimals
130.     pearsoncorrel.append(np.around(np.array(pearsoncorrelation),2))
131.
132. # print the correlation values
133. # for i in range(4):
134. #     print(f'Pearson coefficient of {xlabel[i]} and {ylabel[i]} is {pearsoncorrel[i]}')
135. #     print('\n')
136.
137.
138. # plotting the scatter plot
139.
140. i = 0
141.
142. f, axs = plt.subplots(2, 2, sharey=True, constrained_layout=False)
143. f.suptitle('Correlation replicates', fontsize=16, y = 0.95)
144. f.subplots_adjust(left=None, bottom=None, right=None, top=None, wspace=0.2,
145.                    hspace=0.4)
```

Materials and Methods Chapter 3.3

```
146. for ax in [(0,0),(0,1), (1,0), (1,1)]:
147.     color_list = ['#37b067', '#bbe2ad', '#6296bc', '#b9d2d5', '#edb40e', '#eed39d', '#eb6672',
                    '#eea7a7', '#9f8cae', '#d0cadb', '#7fd7c1', '#bbf2d5', '#376c72', '#7db8ba', '#ee9dcc',
                    '#f1c9dd', '#e3791c', '#ffc097', '#9f765e', '#dac1b1', '#000000']
148. # color_list = ['grey','blue','grey','blue']
149. color=0
150. fitnessA = Normalization(sample_0, sampleA[i])
151. fitnessB = Normalization(sample_0, sampleB[i])
152.
153.
154. for amino in fitnessA:
155.     # plt.plot(fitnessA[amino], fitnessB[amino], "o", color=color_list[color],
    label=amino, alpha=1, markersize=3)
156.     axs[ax].plot(fitnessA[amino], fitnessB[amino], "o", color=color_list[color],
    label=amino, alpha=1, markersize=2.5)
157.     color += 1
158.
159.
160. # adding annotation
161. #         for indexation in range(len(position)):
162. #             if fitnessA[amino] != None and fitnessB[amino] != None:
163. #                 axs[ax].annotate(text=f'{position[indexation]}',
    xy=(fitnessA[amino][indexation], fitnessB[amino][indexation]))
164.
165.     axs[ax].set_xlabel(xlabel[i],fontweight='bold')
166.     axs[ax].set_ylabel(ylabel[i],fontweight='bold')
167.     axs[ax].set_xlim(0,3)
168.     axs[ax].set_ylim(0,3)
169.     axs[ax].text(0.25,2.75,f'R = {pearsoncorrel[i]}', style='normal',weight='bold', fontsize=9,
    color = 'black')#,bbox={'facecolor': 'white', 'alpha': 0.5, 'pad': 5})
170.
171. plt.show()
172. plt.close()
```

- D. For creating the logoplot using the dmslogo package. The source code for dmslogo is available at <https://github.com/jbloomlab/dmslogo>

```
1. import os
2. import dmslogo
3. import pandas as pd
4. import matplotlib.pyplot as plt
5. import dmslogo.utils
6.
7. pd.set_option('display.max_rows', None)
8. # pd.set_option('display.max_columns', None)
```

Materials and methods: Appendices

```
9.  output = "../output/"
10.
11. # loading the csv with enrichment values
12.
13. os.chdir("../data/values_enrichment_logoplot")
14. output_df = pd.read_csv('+DA_output_df_raptorX.csv')
15.
16. j = 0
17. ss_raptor_overlay_list = []
18.
19. for i in range(len(output_df.site)):
20.     if output_df['ss_raptor'][j] == 'C':
21.         ss_raptor_overlay_list.append('#6080ff')
22.     elif output_df['ss_raptor'][j] == 'H':
23.         ss_raptor_overlay_list.append('#f00080')
24.     elif output_df['ss_raptor'][j] == 'E':
25.         ss_raptor_overlay_list.append('#ffff00')
26.     j += 1
27.
28. j = 0
29. rsa_raptor_overlay_list = []
30.
31. for i in range(len(output_df.site)):
32.     if output_df['rsa_raptor'][j] == 'E':
33.         rsa_raptor_overlay_list.append('#001eff')
34.     elif output_df['rsa_raptor'][j] == 'M':
35.         rsa_raptor_overlay_list.append('#ff9900')
36.     elif output_df['rsa_raptor'][j] == 'B':
37.         rsa_raptor_overlay_list.append('#ff0000')
38.     j += 1
39.
40. output_df['ss_raptor_overlay'] = ss_raptor_overlay_list
41. output_df['rsa_raptor_overlay'] = rsa_raptor_overlay_list
42.
43. output_df.to_csv(output+'+DA_output_df_raptorX_withoverlay.csv', index=False)
44.
45.
46. # loading the csv with enrichment values
47. os.chdir("../data/values_enrichment_logoplot")
48. output_df_withDA = pd.read_csv('+DA_output_df_raptorX_withoverlay.csv')
49. output_df_withoutDA = pd.read_csv('-DA_output_df_raptorX_withoverlay.csv')
50.
51. # plotting the graphs
52.
53. # number 1 for with DA
54. fig1, ax1 = dmslogo.draw_logo(data=output_df_withDA,
55.                                x_col='site',
56.                                letter_col='letter',
57.                                letter_height_col='height',
58.                                color_col='color',
```

Materials and Methods Chapter 3.3

```
59.         xlabel='position',
60.         ylabel='enrichment',
61.         title='+DA_combined',
62.         # axisfontscale=2.5,
63.         heatmap_overlays=['ss_raptor_overlay', 'rsa_raptor_overlay'],
64.         # widthscale=4,
65.         heightscale=1.5,
66.     )
67.
68. # plt.savefig(output + f'{file_name}.eps', format='eps')
69. plt.savefig('+DA_raptorx.eps', format='eps', bbox_inches='tight')
70.
71. # number 2 for without DA
72. fig2, ax2 = dmslogo.draw_logo(data=output_df_withoutDA,
73.                               x_col='site',
74.                               letter_col='letter',
75.                               letter_height_col='height',
76.                               color_col='color',
77.                               xlabel='position',
78.                               ylabel='enrichment',
79.                               title='-DA_combined',
80.                               # axisfontscale=2.5,
81.                               heatmap_overlays=['ss_raptor_overlay', 'rsa_raptor_overlay'],
82.                               # widthscale=4,
83.                               heightscale=1.5,
84.                               )
85.
86. # plt.savefig(output + f'{file_name}.eps', format='eps')
87. plt.savefig('-DA_raptorx.eps', format='eps', bbox_inches='tight')
```


5.

Bibliography

- Adkar, Bharat V. et al. 2012. "Protein Model Discrimination Using Mutational Sensitivity Derived from Deep Sequencing." *Structure* 20(2): 371–81.
- Aharoni, Amir et al. 2005. "The 'evolvability' of Promiscuous Protein Functions." *Nature Genetics* 37(1): 73–76.
- Araya, Carlos L. et al. 2012. "A Fundamental Protein Property, Thermodynamic Stability, Revealed Solely from Large-Scale Measurements of Protein Function." *Proceedings of the National Academy of Sciences of the United States of America* 109(42): 16858–63.
- Araya, Carlos L., and Douglas M. Fowler. 2011. "Deep Mutational Scanning: Assessing Protein Function on a Massive Scale." *Trends in Biotechnology* 29(9): 435–42.
- Arjes, Heidi A. et al. 2021. "Three-Dimensional Biofilm Colony Growth Supports a Mutualism Involving Matrix and Nutrient Sharing." *eLife* 10: 1–79.
- Ast, O. et al. 1998. "A Rapid and Sensitive Bacterial Assay to Determine the Inhibitory Effect of 'interface' Peptides on HIV-1 Protease Co-Expressed in Escherichia Coli." *Journal of Virological Methods* 71(1): 77–85.
- Aw, Rochelle, and Karen M. Polizzi. 2013. "Can Too Many Copies Spoil the Broth?" *Microbial Cell Factories* 12(1): 1–9.
<https://link.springer.com/articles/10.1186/1475-2859-12-128> (June 21, 2021).
- Babél, Lilia M., Christopher J. Linneversl, and Brian F. Schmidt. 2000.

- “Production of Active Mammalian and Viral Proteases in Bacterial Expression Systems.” *Biotechnology and Genetic Engineering Reviews* 17(1): 213–54.
- Bailey-Elkin, Ben A. et al. 2014. “Crystal Structure of the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) Papain-like Protease Bound to Ubiquitin Facilitates Targeted Disruption of Deubiquitinating Activity to Demonstrate Its Role in Innate Immune Suppression.” *Journal of Biological Chemistry* 289(50): 34667–82. <http://dx.doi.org/10.1074/jbc.M114.609644>.
- Balint, Robert F., and Ingrid Plooy. 1995. “Protease-Dependent Streptomycin Sensitivity in *E. Coli*—a System for Protease Inhibitor Selection.” *Bio/Technology* 13(5): 507–10.
- Bershtein, Shimon, Korina Goldin, and Dan S. Tawfik. 2008. “Intense Neutral Drifts Yield Robust and Evolvable Consensus Proteins.” *Journal of Molecular Biology* 379(5): 1029–44.
- Block, T. M., and R. H. Grafstrom. 1990. “Novel Bacteriological Assay for Detection of Potential Antiviral Agents.” *Antimicrobial Agents and Chemotherapy* 34(12): 2337–41.
- Blomqvist, Trinelise, Hilde Steinmoen, and Leiv Sigve Håvarstein. 2006. “Natural Genetic Transformation: A Novel Tool for Efficient Genetic Engineering of the Dairy Bacterium *Streptococcus Thermophilus*.” *Applied and Environmental Microbiology* 72(10): 6751–56.
- Bolotin, Alexander et al. 2004. “Complete Sequence and Comparative Genome Analysis of the Dairy Bacterium *Streptococcus Thermophilus*.” *Nature Biotechnology* 22(12): 1554–58.
- Bordoli, Lorenza, and Torsten Schwede. 2012. “Automated Protein Structure Modeling with Swiss-Model Workspace and the Protein Model Portal.” *Methods in Molecular Biology* 857: 107–36.
- Bornberg-Bauer, E., and H. S. Chan. 1999. “Modeling Evolutionary Landscapes: Mutational Stability, Topology, and Superfunnels in Sequence Space.” *Proceedings of the National Academy of Sciences* 96(19): 10689–94. <http://www.pnas.org/cgi/doi/10.1073/pnas.96.19.10689>.
- Boucher, Jeffrey I. et al. 2014. “Viewing Protein Fitness Landscapes through a Next-Gen Lens.” *Genetics* 198(2): 461–71.

Bibliography Chapter 3.3

- Brennan, Michael J. 2017. "Biofilms and Mycobacterium Tuberculosis" ed. Sabine Ehrt. *Infection and Immunity* 85(10): 2017. <https://journals.asm.org/doi/10.1128/IAI.00411-17>.
- Brown, Celeste J. et al. 2002. "Evolutionary Rate Heterogeneity in Proteins with Long Disordered Regions." *Journal of Molecular Evolution* 55(1): 104–10. <https://pubmed.ncbi.nlm.nih.gov/12165847/> (June 21, 2021).
- Bustamante, Carlos D., Jeffrey P. Townsend, and Daniel L. Hartl. 2000. "Solvent Accessibility and Purifying Selection within Proteins of Escherichia Coli and Salmonella Enterica." *Molecular Biology and Evolution* 17(2): 301–8. <https://pubmed.ncbi.nlm.nih.gov/10677853/> (June 21, 2021).
- Cáceres, Nancy E. et al. 1997. "Overexpression of the D-Alanine Racemase Gene Confers Resistance to D- Cycloserine in Mycobacterium Smegmatis." *Journal of Bacteriology* 179(16): 5046–55.
- Caminero, José A., Giovanni Sotgiu, Alimuddin Zumla, and Giovanni Battista Migliori. 2010. "Best Drug Treatment for Multidrug-Resistant and Extensively Drug-Resistant Tuberculosis." *The Lancet Infectious Diseases* 10(9): 621–29. [http://dx.doi.org/10.1016/S1473-3099\(10\)70139-0](http://dx.doi.org/10.1016/S1473-3099(10)70139-0).
- Chen, Jiazhen et al. 2017. "Identification of Novel Mutations Associated with Cycloserine Resistance in Mycobacterium Tuberculosis." *Journal of Antimicrobial Chemotherapy* 72(12): 3272–76.
- Cheng, Lin et al. 2007. "Discovery of Antibacterial Cyclic Peptides That Inhibit the ClpXP Protease." *Protein Science* 16(8): 1535–42.
- de Chiara, Cesira et al. 2020. "D-Cycloserine Destruction by Alanine Racemase and the Limit of Irreversible Inhibition." *Nature Chemical Biology* 16(6): 686–94.
- Choi, Yongwook et al. 2012. "Predicting the Functional Effect of Amino Acid Substitutions and Indels." *PLoS ONE* 7(10).
- Creager, Angela N. H. 2001. "Hershey Heaven." *Nature Structural Biology* 8(1): 18–19. <http://www.nature.com/doi/10.1038/82991> (June 11, 2021).
- Dandage, Rohan et al. 2018. "Differential Strengths of Molecular

- Determinants Guide Environment Specific Mutational Fates." *PLoS Genetics* 14(5): e1007419.
<https://doi.org/10.1371/journal.pgen.1007419> (May 6, 2021).
- Dandoy, Damien et al. 2011. "The Fast Milk Acidifying Phenotype of *Streptococcus Thermophilus* Can Be Acquired by Natural Transformation of the Genomic Island Encoding the Cell-Envelope Proteinase PrtS." *Microbial Cell Factories* 10(Suppl 1): S21.
<https://pubmed.ncbi.nlm.nih.gov/21995822/> (June 9, 2021).
- Dautin, N., G. Karimova, A. Ullmann, and D. Ladant. 2000. "Sensitive Genetic Screen for Protease Activity Based on a Cyclic AMP Signaling Cascade in *Escherichia Coli*." *Journal of Bacteriology* 182(24): 7060–66.
- Dean, Antony M., Claudia Neuhauser, Elise Grenier, and G. Brian Golding. 2002. "The Pattern of Amino Acid Replacements in α/β -Barrels." *Molecular Biology and Evolution* 19(11): 1846–64.
<https://academic.oup.com/mbe/article/19/11/1846/1012440> (June 21, 2021).
- Defoor, Els, Maj Britt Kryger, and Jan Martinussen. 2007. "The Orotate Transporter Encoded by OroP from *Lactococcus Lactis* Is Required for Orotate Utilization and Has Utility as a Food-Grade Selectable Marker." *Microbiology* 153(11): 3645–59.
- Delcour, Jean et al. 1999. 76 Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology *The Biosynthesis and Functionality of the Cell-Wall of Lactic Acid Bacteria*.
- Desjardins, Christopher A. et al. 2016. "Genomic and Functional Analyses of *Mycobacterium Tuberculosis* Strains Implicate Ald in D-Cycloserine Resistance." *Nature Genetics* 48(5): 544–51.
- Devaraj, Santhana G. et al. 2007. "Regulation of IRF-3-Dependent Innate Immunity by the Papain-like Protease Domain of the Severe Acute Respiratory Syndrome Coronavirus." *Journal of Biological Chemistry* 282(44): 32208–21. <http://dx.doi.org/10.1074/jbc.M704870200>.
- Dorrazehi, Gol Mohammad et al. 2020. "Building Scarless Gene Libraries in the Chromosome of Bacteria." In *Peptide and Protein Engineering*, , 189–211. http://link.springer.com/10.1007/978-1-0716-0720-6_11.

Bibliography Chapter 3.3

- Drummond, D. Allan et al. 2005. "Why Highly Expressed Proteins Evolve Slowly." *Proceedings of the National Academy of Sciences of the United States of America* 102(40): 14338–43.
- Echave, Julian, Stephanie J. Spielman, and Claus O. Wilke. 2016. "Causes of Evolutionary Rate Variation among Protein Sites." *Nature Reviews Genetics* 17(2): 109–21.
<http://dx.doi.org/10.1038/nrg.2015.18>.
- Ernst, Andreas et al. 2010. "Coevolution of PDZ Domain-Ligand Interactions Analyzed by High-Throughput Phage Display and Deep Sequencing." *Molecular BioSystems* 6(10): 1782–90.
- Esaki, Nobuyoshi, and Christopher T. Walsh. 1986. "Biosynthetic Alanine Racemase of Salmonella Typhimurium: Purification and Characterization of the Enzyme Encoded by the Air Gene." *Biochemistry* 25(11): 3261–67.
<https://pubmed.ncbi.nlm.nih.gov/3524677/> (June 4, 2021).
- Evangelopoulos, Dimitrios et al. 2019. "Comparative Fitness Analysis of D-Cycloserine Resistant Mutants Reveals Both Fitness-Neutral and High-Fitness Cost Genotypes." *Nature Communications*.
- Fenn, Timothy D., Geoffrey F. Stamper, Anthony A. Morollo, and Dagmar Ringe. 2003. "A Side Reaction of Alanine Racemase: Transamination of Cycloserine." *Biochemistry* 42(19): 5775–83.
- Fernandez-Espla, M. D., P. Garault, V. Monnet, and F. Rul. 2000. "Streptococcus Thermophilus Cell Wall-Anchored Proteinase: Release, Purification, and Biochemical and Genetic Characterization." *Applied and Environmental Microbiology* 66(11): 4772–78. <https://pubmed.ncbi.nlm.nih.gov/11055922/> (June 22, 2021).
- Firth, Andrew E., and Wayne M. Patrick. 2005. "Statistics of Protein Library Construction." *Bioinformatics* 21(15): 3314–15.
- Fontaine, Laetitia, Céline Boutry, Marie Henry De Frahan, Brigitte Delplace, Christophe Fremaux, et al. 2010. "A Novel Pheromone Quorum-Sensing System Controls the Development of Natural Competence in Streptococcus Thermophilus and Streptococcus Salivarius †." *JOURNAL OF BACTERIOLOGY* 192(5): 1444–54.
<http://jb.asm.org/> (April 22, 2019).

- Fontaine, Laetitia, Damien Dandoy, Céline Boutry, Brigitte Delplace, Marie Henry De Frahan, et al. 2010. "Development of a Versatile Procedure Based on Natural Transformation for Marker-Free Targeted Genetic Modification in *Streptococcus Thermophilus*." *Applied and Environmental Microbiology* 76(23): 7870–77. <http://www.ncbi.nlm.nih.gov/pubmed/20935129> (April 23, 2019).
- Fontaine, Laetitia et al. 2013a. "Mechanism of Competence Activation by the ComRS Signalling System in *Streptococci*." *Molecular Microbiology* 87(6): 1113–32.
- Forsyth, Charles M. et al. 2013. "Deep Mutational Scanning of an Antibody against Epidermal Growth Factor Receptor Using Mammalian Cell Display and Massively Parallel Pyrosequencing." *mAbs* 5(4): 523–32. <https://pubmed.ncbi.nlm.nih.gov/23765106/> (June 10, 2021).
- Fowler, Douglas M., Carlos L. Araya, Wayne Gerard, and Stanley Fields. 2011. "Enrich: Software for Analysis of Protein Function by Enrichment and Depletion of Variants." *Bioinformatics* 27(24): 3430–31.
- Fowler, Douglas M., and Stanley Fields. 2014. "Deep Mutational Scanning: A New Style of Protein Science." *Nature Methods* 11(8): 801–7.
- Franzosa, Eric A., and Yu Xia. 2009. "Structural Determinants of Protein Evolution Are Context-Sensitive at the Residue Level." *Molecular Biology and Evolution* 26(10): 2387–95.
- Frieman, Matthew et al. 2009. "Severe Acute Respiratory Syndrome Coronavirus Papain-Like Protease Ubiquitin-Like Domain and Catalytic Domain Regulate Antagonism of IRF3 and NF-KB Signaling." *Journal of Virology* 83(13): 6689–6705.
- Fujino, Yasuhiro et al. 2012. "Robust in Vitro Affinity Maturation Strategy Based on Interface-Focused High-Throughput Mutational Scanning." *Biochemical and Biophysical Research Communications* 428(3): 395–400. <https://pubmed.ncbi.nlm.nih.gov/23103372/> (June 10, 2021).
- Galka, Pierre, Elisabeth Jamez, Gilles Joachim, and Patrice Soumillion. 2017. "QuickLib, a Method for Building Fully Synthetic Plasmid Libraries by Seamless Cloning of Degenerate Oligonucleotides."

Bibliography Chapter 3.3

- PLoS ONE* 12(4).
- Gardan, Rozenn et al. 2009. "The Oligopeptide Transport System Is Essential for the Development of Natural Competence in *Streptococcus Thermophilus* Strain LMD-9." *Journal of Bacteriology* 191(14): 4647–55.
- Geiler-Samerotte, Kerry A. et al. 2011. "Misfolded Proteins Impose a Dosage-Dependent Fitness Cost and Trigger a Cytosolic Unfolded Protein Response in Yeast." *Proceedings of the National Academy of Sciences of the United States of America* 108(2): 680–85.
- Glaser, Fabian et al. 2005. "The ConSurf-HSSP Database: The Mapping of Evolutionary Conservation among Homologs onto PDB Structures." *Proteins: Structure, Function and Genetics* 58(3): 610–17.
<https://www.onlinelibrary.wiley.com/doi/full/10.1002/prot.20305>
- Gold, Matthew G. et al. 2006. "Molecular Basis of AKAP Specificity for PKA Regulatory Subunits." *Molecular Cell* 24(3): 383–95.
- Goldman, Nick, Jeffrey L. Thorne, and David T. Jones. 1998. "Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution." *Genetics* 149(1): 445–58.
[/pmc/articles/PMC1460119/?report=abstract](https://pmc/articles/PMC1460119/?report=abstract) (June 21, 2021).
- Gregoret, L M, and R T Sauer. 1993. "Additivity of Mutant Effects Assessed by Binomial Mutagenesis." *Proceedings of the National Academy of Sciences* 90(9): 4246–50.
<https://www.pnas.org/content/90/9/4246> (August 9, 2021).
- Grimm, Bernhard, Marvin A. Smith, and Diter von Wettstein. 1992. "The Role of Lys272 in the Pyridoxal 5-phosphate Active Site of *Synechococcus* Glutamate-1-semialdehyde Aminotransferase." *European Journal of Biochemistry* 206(2): 579–85.
- Haddox, Hugh K. et al. 2018. "Mapping Mutational Effects along the Evolutionary Landscape of HIV Envelope." *eLife* 7: 1–29.
- Hall, Kelsi R. et al. 2020. "Intracellular Complexities of Acquiring a New Enzymatic Function Revealed by Mass-Randomisation of Active Site Residues." *eLife* 9: 1–40.
- Haustenne, Laurie, Georges Bastin, Pascal Hols, and Laetitia Fontaine. 2015. "Modeling of the ComRS Signaling Pathway Reveals the Limiting Factors Controlling Competence in *Streptococcus*

- Thermophilus." *Frontiers in Microbiology* 6(DEC).
- Hietpas, R. T., J. D. Jensen, and D. N. A. Bolon. 2011. "Experimental Illumination of a Fitness Landscape." *Proceedings of the National Academy of Sciences* 108(19): 7896–7901.
<http://www.pnas.org/cgi/doi/10.1073/pnas.1016024108>.
- Hoffmann, Kai, Elisabeth Schneider-Scherzer, Horst Kleinkauf, and Rainer Zocher. 1994. "Purification and Characterization of Eucaryotic Alanine Racemase Acting as Key Enzyme in Cyclosporin Biosynthesis." *Journal of Biological Chemistry* 269(17): 12710–14.
- Ribeiro, António J.M. et al. 2018. "Mechanism and Catalytic Site Atlas (M-CSA): A Database of Enzyme Reaction Mechanisms and Active Sites." *Nucleic Acids Research* 46(D1): D618–23.
- Hols, Pascal, Frédéric Hancy, Laetitia Fontaine, Benoît Grossiord, Deborah Prozzi, Nathalie Leblond-Bourget, Bernard Decaris, Alexander Bolotin, Christine Delorme, S. Dusko Ehrlich, et al. 2005. "New Insights in the Molecular Biology and Physiology of Streptococcus Thermophilus Revealed by Comparative Genomics." *FEMS Microbiology Reviews* 29(3 SPEC. ISS.): 435–63.
<http://doi.wiley.com/10.1016/j.femsre.2005.04.008>.
- Huang, He, Jovan Damjanovic, Jiayuan Miao, and Yu-Shan Lin. 2021. "Cyclic Peptides: Backbone Rigidification and Capability of Mimicking Motifs at Protein–Protein Interfaces." *Physical Chemistry Chemical Physics* 23(1): 607–16.
<http://xlink.rsc.org/?DOI=D0CP04633G>.
- Huang, Tsun Tsao, María Laura Del Valle Marcos, Jenn Kang Hwang, and Julian Echave. 2014. "A Mechanistic Stress Model of Protein Evolution Accounts for Site-Specific Evolutionary Rates and Their Relationship with Packing Density and Flexibility." *BMC Evolutionary Biology* 14(1): 1–10. <http://www.biomedcentral.com/1471-2148/14/78> (June 21, 2021).
- Im, Hookang et al. 2011. "The Crystal Structure of Alanine Racemase from Streptococcus Pneumoniae, a Target for Structure-Based Drug Design." *BMC Microbiology* 11: 1–15.
- Inagaki, Kenji et al. 2000. "Purification and Some Characteristics of a Monomeric Alanine Racemase from an Extreme Thermophile,

Bibliography Chapter 3.3

- Thermus Thermophilus.” *Journal of Bioscience and Bioengineering* 90(3): 344–46.
- Jack, Benjamin R., Austin G. Meyer, Julian Echave, and Claus O. Wilke. 2016. “Functional Sites Induce Long-Range Evolutionary Constraints in Enzymes.” *PLoS Biology*.
- James, L. C., Tawfik, Dan S. 2001. “Catalytic and Binding Poly-Reactivities Shared by Two Unrelated Proteins: The Potential Role of Promiscuity in Enzyme Evolution.” *Protein Science* 10(12): 2600–2607. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2374036/>.
- Jiang, Li et al. 2013. “Latent Effects of Hsp90 Mutants Revealed at Reduced Expression Levels” ed. Tricia R. Serio. *PLoS Genetics* 9(6): e1003600. www.plosgenetics.org (May 2, 2019).
- Johnson, L. Steven, Sean R. Eddy, and Elon Portugaly. 2010. “Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure.” *BMC Bioinformatics* 11(1): 1–8. <http://ekhidna.biocenter>. (June 18, 2021).
- Jones, Eric M. et al. 2020. “Structural and Functional Characterization of G Protein-Coupled Receptors with Deep Mutational Scanning.” *eLife* 9: 1–52.
- Kaltenbach, Miriam, Stephane Emond, Florian Hollfelder, and Nobuhiko Tokuriki. 2016. “Functional Trade-Offs in Promiscuous Enzymes Cannot Be Explained by Intrinsic Mutational Robustness of the Native Activity.” *PLoS Genetics* 12(10): 1–18.
- Kanehisa, Minoru, and Susumu Goto. 2000. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 28(1): 27–30. <https://pubmed.ncbi.nlm.nih.gov/10592173/> (June 4, 2021).
- Kang, Lishan et al. 2011. “Upregulation of MetC Is Essential for D-Alanine-Independent Growth of an Alr/DadX-Deficient Escherichia Coli Strain.” *Journal of Bacteriology* 193(5): 1098–1106. <https://journals.asm.org/journal/jb> (June 4, 2021).
- Katsube, Satoshi et al. 2016. “Secretion of D-Alanine by Escherichia Coli.” *Microbiology (United Kingdom)* 162(7): 1243–52.
- Khanal, Akhil, Sean Yu McLoughlin, Jamie P. Kershner, and Shelley D. Copley. 2015. “Differential Effects of a Mutation on the Normal and Promiscuous Activities of Orthologs: Implications for Natural and

- Directed Evolution.” *Molecular Biology and Evolution* 32(1): 100–108.
- Khersonsky, Olga, and Dan S. Tawfik. 2010. “Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective.” *Annual Review of Biochemistry* 79: 471–505.
- Kim, Dae Kyum et al. 2020. “A Comprehensive, Flexible Collection of SARS-CoV-2 Coding Regions.” *G3: Genes, Genomes, Genetics* 10(9).
- Kim, Ikjin, Christina R. Miller, David L. Young, and Stanley Fields. 2013. “High-Throughput Analysis of in Vivo Protein Stability.” *Molecular and Cellular Proteomics* 12(11): 3370–78.
<https://pubmed.ncbi.nlm.nih.gov/23897579/> (June 10, 2021).
- Kim, Philip M., Long J. Lu, Yu Xia, and Mark B. Gerstein. 2006. “Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights.” *Science* 314(5807): 1938–41.
<https://science.sciencemag.org/content/314/5807/1938> (June 21, 2021).
- Kjos, Morten. 2019. “Transcriptional Knockdown in Pneumococci Using CRISPR Interference.” In *Methods in Molecular Biology*, , 89–98.
http://link.springer.com/10.1007/978-1-4939-9199-0_8.
- Klesmith, Justin R. et al. 2017. “Trade-Offs between Enzyme Fitness and Solubility Illuminated by Deep Mutational Scanning.” *Proceedings of the National Academy of Sciences of the United States of America* 114(9): 2265–70.
- Krissinel, Evgeny, and Kim Henrick. 2007. “Inference of Macromolecular Assemblies from Crystalline State.” *Journal of Molecular Biology* 372(3): 774–97. <https://pubmed.ncbi.nlm.nih.gov/17681537/> (June 6, 2021).
- Kulka, Kathleen, Graham Hatfull, and Anil K. Ojha. 2012. “Growth of Mycobacterium Tuberculosis Biofilms.” *Journal of Visualized Experiments* (60): e3820. <https://www.jove.com/t/3820>.
- Kulshreshtha, Shweta, Vigi Chaudhary, Girish K. Goswami, and Nidhi Mathur. 2016. “Computational Approaches for Predicting Mutant Protein Stability.” *Journal of Computer-Aided Molecular Design* 30(5): 401–12.
- Kumarasinghe, Isuru R., and Patrick M. Woster. 2014. “Synthesis and

Bibliography Chapter 3.3

- Evaluation of Novel Cyclic Peptide Inhibitors of Lysine-Specific Demethylase 1." *ACS Medicinal Chemistry Letters* 5(1): 29–33. <https://pubs.acs.org/sharingguidelines> (June 19, 2021).
- Lambert, Mary Pulliam, and Francis C. Neuhaus. 1972. "Factors Affecting the Level of Alanine Racemase in Escherichia Coli1." *Journal of Bacteriology* 109(3): 1156–61. <https://jlb.asm.org/content/109/3/1156>.
- Lecomte, Xavier et al. 2016. "Streptococcus Thermophilus, an Emerging and Promising Tool for Heterologous Expression: Advantages and Future Trends." *Food Microbiology* 53: 2–9.
- LeMagueres, Pierre et al. 2003. "Crystal Structure at 1.45 Å Resolution of Alanine Racemase from a Pathogenic Bacterium, Pseudomonas Aeruginosa, Contains Both Internal and External Aldimine Forms." *Biochemistry* 42(50): 14752–61.
- . 2005a. "The 1.9 Å Crystal Structure of Alanine Racemase from Mycobacterium Tuberculosis Contains a Conserved Entryway into the Active Site." *Biochemistry* 44(5): 1471–81.
- Letort, C., and V. Juillard. 2001. "Development of a Minimal Chemically-Defined Medium for the Exponential Growth of Streptococcus Thermophilus." *Journal of Applied Microbiology* 91(6): 1023–29.
- Lobocka, M., J. Hennig, J. Wild, and T. Klopotoski. 1994. "Organization and Expression of the Escherichia Coli K-12 Dad Operon Encoding the Smaller Subunit of D-Amino Acid Dehydrogenase and the Catabolic Alanine Racemase." *Journal of Bacteriology* 176(5): 1500–1510. <https://pubmed.ncbi.nlm.nih.gov/7906689/> (June 4, 2021).
- Major, Dan T., Kwangho Nam, and Jiali Gao. 2006. "Transition State Stabilization and α -Amino Carbon Acidity in Alanine Racemase." *Journal of the American Chemical Society* 128(25): 8114–15.
- Major, Dan Thomas, and Jiali Gao. 2006. "A Combined Quantum Mechanical and Molecular Mechanical Study of the Reaction Mechanism and α -Amino Acidity in Alanine Racemase." *Journal of the American Chemical Society* 128(50): 16345–57.
- Marcos, María Laura, and Julian Echave. 2015. "Too Packed to Change: Side-Chain Packing and Site-Specific Substitution Rates in Protein Evolution." *PeerJ* 2015(3): e911. <https://peerj.com/articles/911>

(June 11, 2021).

- Martinussen, J., J. Schallert, B. Andersen, and K. Hammer. 2001. "The Pyrimidine Operon PyrRPB-CarA from *Lactococcus Lactis*." *Journal of Bacteriology* 183(9): 2785–94. <http://jb.asm.org/> (April 25, 2019).
- Mate, Diana, Eva Garcia-Ruiz, Susana Camarero, and Miguel Alcalde. 2011. "Directed Evolution of Fungal Laccases." *Current Genomics* 12(2): 113–22. [/pmc/articles/PMC3129045/](https://pubmed.ncbi.nlm.nih.gov/2129045/) (June 9, 2021).
- Mazurenko, Stanislav. 2020. "Predicting Protein Stability and Solubility Changes upon Mutations: Data Perspective." *ChemCatChem* 12(22): 5590–98.
- McCall, J. Owen, Sunil Kadam, and Leonard Katz. 1994. "A High Capacity Microbial Screen for Inhibitors of Human Rhinovirus Protease 3C." *Bio/Technology* 12(10): 1012–16. <http://www.nature.com/doi/10.1038/nbt1094-1012>.
- McLaughlin, Richard N. et al. 2012. "The Spatial Architecture of Protein Function and Adaptation." *Nature* 491(7422): 138–42.
- McLoughlin, Sean Yu, and Shelley D. Copley. 2008. "A Compromise Required by Gene Sharing Enables Survival: Implications for Evolution of New Enzyme Activities." *Proceedings of the National Academy of Sciences of the United States of America* 105(36): 13497–502.
- Mehlhoff, Jacob D. et al. 2020. "Collateral Fitness Effects of Mutations." *Proceedings of the National Academy of Sciences of the United States of America* 117(21).
- Melamed, Daniel et al. 2013. "Deep Mutational Scanning of an RRM Domain of the *Saccharomyces Cerevisiae* Poly(A)-Binding Protein." *RNA* 19(11): 1537–51. <https://pubmed.ncbi.nlm.nih.gov/24064791/> (June 10, 2021).
- Mell, Joshua Chang, and Rosemary J. Redfield. 2014. "Natural Competence and the Evolution of DNA Uptake Specificity." *Journal of Bacteriology* 196(8): 1471–83.
- Mignolet, Johann et al. 2018. "Circuitry Rewiring Directly Couples Competence to Predation in the Gut Dweller *Streptococcus Salivarius*." *Cell Reports* 22(7): 1627–38. <http://linkinghub.elsevier.com/retrieve/pii/S2211124718301049>.

Bibliography Chapter 3.3

- Mikkelsen, Tarjei S. et al. 2007. "Genome-Wide Maps of Chromatin State in Pluripotent and Lineage-Committed Cells." *Nature* 448(7153): 553–60. <https://www.nature.com/articles/nature06008> (June 10, 2021).
- Milligan, Daniel L. et al. 2007. "The Alanine Racemase of *Mycobacterium Smegmatis* Is Essential for Growth in the Absence of D-Alanine." *Journal of Bacteriology* 189(22): 8381–86.
- Mintseris, Julian, and Zhiping Weng. 2005. "Structure, Function, and Evolution of Transient and Obligate Protein-Protein Interactions." *Proceedings of the National Academy of Sciences of the United States of America* 102(31): 10930–35.
www.pnas.org/cgi/doi/10.1073/pnas.0502667102 (June 21, 2021).
- Mistry, Jaina et al. 2021. "Pfam: The Protein Families Database in 2021." *Nucleic Acids Research* 49(D1): D412–19.
<https://academic.oup.com/nar/article/49/D1/D412/5943818> (June 6, 2021).
- Morollo, Anthony A., Gregory A. Petsko, and Dagmar Ringe. 1999. "Structure of a Michaelis Complex Analogue: Propionate Binds in the Substrate Carboxylate Site of Alanine Racemase." *Biochemistry* 38(11): 3293–3301.
<https://pubs.acs.org/doi/abs/10.1021/bi9822729> (June 6, 2021).
- Mu, Yingli, Yongping Xin, Tingting Guo, and Jian Kong. 2020. "Identification and Characterization of a Moonlighting Protein-Enolase for Surface Display in *Streptococcus Thermophilus*." *Microbial Cell Factories* 19(1): 1–10.
<https://doi.org/10.1186/s12934-020-01389-y> (June 14, 2021).
- Mustata, Gabriela, and James M. Briggs. 2004. "Cluster Analysis of Water Molecules in Alanine Racemase and Their Putative Structural Role." *Protein Engineering, Design and Selection* 17(3): 223–34.
- Nagalakshmi, Ugrappa et al. 2008. "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing." *Science* 320(5881): 1344–49. <https://pubmed.ncbi.nlm.nih.gov/18451266/> (June 10, 2021).
- Nakatani, Yoshio et al. 2017. "Role of Alanine Racemase Mutations in *Mycobacterium Tuberculosis* D-Cycloserine Resistance."

Antimicrobial Agents and Chemotherapy.

- Naumann, Todd A., Ali Tavassoli, and Stephen J. Benkovic. 2008. "Genetic Selection of Cyclic Peptide Dam Methyltransferase Inhibitors." *ChemBioChem* 9(2): 194–97.
- Nevin Gerek, Zeynep, Sudhir Kumar, and Sefika Banu Ozkan. 2013. "Structural Dynamics Flexibility Informs Function and Evolution at a Proteome Scale." *Evolutionary Applications* 6(3): 423–33. [/pmc/articles/PMC3673471/](https://pmc/articles/PMC3673471/) (June 21, 2021).
- Newton, Matilda S., Vickery L. Arcus, and Wayne M. Patrick. 2015. "Rapid Bursts and Slow Declines: On the Possible Evolutionary Trajectories of Enzymes." *Journal of the Royal Society Interface* 12(107).
- Ng, Pauline C., and Steven Henikoff. 2001. "Predicting Deleterious Amino Acid Substitutions." *Genome Research* 11(5): 863–74.
- Ng, Sarah B. et al. 2009. "Targeted Capture and Massively Parallel Sequencing of 12 Human Exomes." *Nature* 461(7261): 272–76. <https://www.nature.com/articles/nature08250> (June 10, 2021).
- . 2010. "Exome Sequencing Identifies MLL2 Mutations as a Cause of Kabuki Syndrome." *Nature Genetics* 42(9): 790–93. <https://pubmed.ncbi.nlm.nih.gov/20711175/> (June 10, 2021).
- Noda-García, Lianet et al. 2019. "Chance and Pleiotropy Dominate Genetic Diversity in Complex Bacterial Environments." *Nature Microbiology* 4(7): 1221–30. <http://dx.doi.org/10.1038/s41564-019-0412-y>.
- Noda, Masafumi, Yasuyuki Matoba, Takanori Kumagai, and Masanori Sugiyama. 2004. "Structural Evidence That Alanine Racemase from a D-Cycloserine-Producing Microorganism Exhibits Resistance to Its Own Product." *Journal of Biological Chemistry* 279(44): 46153–61. <https://pubmed.ncbi.nlm.nih.gov/15302886/> (June 4, 2021).
- Noël, S, and I Sermet-Gaudelus. 2020. "Mucoviscidosis: Fisiopatología, Genética, Aspectos Clínicos y Terapéuticos." *EMC - Pediatría* 55(1 PG-1–23): 1–23. <http://www.sciencedirect.com/science/article/pii/S1245178920434274> NS -.
- Ohta, Tomoko. 1973. "Slightly Deleterious Mutant Substitutions in

Bibliography Chapter 3.3

- Evolution." *Nature* 246(5428): 96–98.
<https://www.nature.com/articles/246096a0> (June 11, 2021).
- Ondrechen, M. J., J. M. Briggs, and J. A. McCammon. 2001. "A Model for Enzyme-Substrate Interaction in Alanine Racemase." *Journal of the American Chemical Society* 123(12): 2830–34.
- Patrick, Wayne M., Andrew E. Firth, and Jonathan M. Blackburn. 2003. "User-Friendly Algorithms for Estimating Completeness and Diversity in Randomized Protein-Encoding Libraries." *Protein Engineering* 16(6): 451–57.
- Patrick, Wayne M., Jan Weisner, and Jonathan M. Blackburn. 2002. "Site-Directed Mutagenesis of Tyr354 in *Geobacillus Stearothermophilus* Alanine Racemase Identifies a Role in Controlling Substrate Specificity and a Possible Role in the Evolution of Antibiotic Resistance." *ChemBioChem* 3(8): 789–92.
- Picard, Didier et al. 1990. "Reduced Levels of Hsp90 Compromise Steroid Receptor Action in Vivo." *Nature* 348(6297): 166–68.
<https://www.nature.com/articles/348166a0> (June 11, 2021).
- Procko, Erik et al. 2013. "Computational Design of a Protein-Based Enzyme Inhibitor." *Journal of Molecular Biology* 425(18): 3563–75.
<https://pubmed.ncbi.nlm.nih.gov/23827138/> (June 10, 2021).
- Ramsey, Duncan C., Michael P. Scherrer, Tong Zhou, and Claus O. Wilke. 2011. "The Relationship between Relative Solvent Accessibility and Evolutionary Rate in Protein Evolution." *Genetics* 188(2): 479–88.
<https://pubmed.ncbi.nlm.nih.gov/21467571/> (June 21, 2021).
- Richards, Jacob P. et al. 2019. "Adaptation of *Mycobacterium Tuberculosis* to Biofilm Growth Is Genetically Linked to Drug Tolerance." *Antimicrobial Agents and Chemotherapy* 63(11): e01213-19..
- Rockah-Shmuel, Liat, Ágnes Tóth-Petróczy, and Dan S. Tawfik. 2015. "Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations." *PLoS Computational Biology* 11(8): 1–28.
- Romero, Philip A., Tuan M. Tran, and Adam R. Abate. 2015. "Dissecting Enzyme Function with Microfluidic-Based Deep Mutational Scanning." *Proceedings of the National Academy of Sciences*.

- Roscoe, Benjamin P. et al. 2013. "Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate." *Journal of Molecular Biology* 425(8): 1363–77.
- Ryan, Owen W. et al. 2014. "Selection of Chromosomal DNA Libraries Using a Multiplex CRISPR System." *eLife* 3(August2014): 1–15.
- di Salvo, Martino L. et al. 2013. "Alanine Racemase from *Tolypocladium Inflatum*: A Key PLP-Dependent Enzyme in Cyclosporin Biosynthesis and a Model of Catalytic Promiscuity." *Archives of Biochemistry and Biophysics* 529(2): 55–65.
<http://dx.doi.org/10.1016/j.abb.2012.11.011>.
- Satoh, Takashi et al. 1996. "Synthetic Peptides Derived from the Fourth Domain of CD4 Antagonize CD4 Function and Inhibit T Cell Activation (D1-D4) Which Show Homology to Members of the Immunoglobulin Superfamily (1). In Particu- CD4 Binds to Nonpolymorphic Regions of the Major H." 443(224): 438–43.
- Schlinkmann, Karola M. et al. 2012. "Critical Features for Biosynthesis, Stability, and Functionality of a G Protein-Coupled Receptor Uncovered by All-versus-All Mutations." *Proceedings of the National Academy of Sciences of the United States of America* 109(25): 9810–15. <https://www.pnas.org/content/109/25/9810> (June 10, 2021).
- Schmidt-Dannert, Claudia, and Frances H Arnold. 1999. "Directed Evolution of Industrial Enzymes." *Trends in Biotechnology* 17(4): 135–36.
<https://linkinghub.elsevier.com/retrieve/pii/S0167779998012839> (August 23, 2019).
- Scott, Charles P. et al. 1999. "Production of Cyclic Peptides and Proteins in Vivo." *Proceedings of the National Academy of Sciences of the United States of America* 96(24): 13638–43.
- Seffernick, J. L., M. L. De Souza, M. J. Sadowsky, and L. P. Wackett. 2001. "Melamine Deaminase and Atrazine Chlorohydrolase: 98 Percent Identical but Functionally Different." *Journal of Bacteriology* 183(8): 2405–10.
- Shahmoradi, Amir et al. 2014. "Predicting Evolutionary Site Variability from Structure in Viral Proteins: Buriedness, Packing, Flexibility, and Design." *Journal of Molecular Evolution* 79(3–4): 130–42.

Bibliography Chapter 3.3

- Shaw, Jeffrey P., Gregory A. Petsko, and Dagmar Ringe. 1997. "Determination of the Structure of Alanine Racemase from *Bacillus Stearothermophilus* at 1.9-Å Resolution ,." *Biochemistry* 36(6): 1329–42. <https://pubs.acs.org/doi/10.1021/bi961856c1/bi961856c> (June 4, 2021).
- Shevchuk, N. A. 2004. "Construction of Long DNA Molecules Using Long PCR-Based Fusion of Several Fragments Simultaneously." *Nucleic Acids Research* 32(2): 19e – 19. <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gnh014>.
- Shih, Chien Hua et al. 2012. "Evolutionary Information Hidden in a Single Protein Structure." *Proteins: Structure, Function and Bioinformatics* 80(6): 1647–57.
- Shin, Donghyuk et al. 2020. "Papain-like Protease Regulates SARS-CoV-2 Viral Spread and Innate Immunity." *Nature* 587(7835): 657–62.
- Sidiq, Karzan R., Man W. Chow, Zhao Zhao, and Richard A. Daniel. 2020. "Alanine Metabolism in *Bacillus Subtilis*." *Molecular Microbiology* (March 2020): 739–57.
- Sinha, Sunita, and Rosemary J. Redfield. 2012. "Natural DNA Uptake by *Escherichia Coli*." *PLoS ONE* 7(4).
- Solem, Christian, Eis Defoor, Peter Ruhdal Jensen, and Jan Martinussen. 2008. "Plasmid PCS1966, a New Selection/Counterselection Tool for Lactic Acid Bacterium Strain Construction Based on the OroP Gene, Encoding an Orotate Transporter from *Lactococcus Lactis*." *Applied and Environmental Microbiology* 74(15): 4772–75.
- Soo, Valerie W.C., Paulina Hanson-Manful, and Wayne M. Patrick. 2011. "Artificial Gene Amplification Reveals an Abundance of Promiscuous Resistance Determinants in *Escherichia Coli*." *Proceedings of the National Academy of Sciences of the United States of America*.
- Soo, Valerie W.C., Yuliana Yosaatmadja, Christopher J. Squire, and Wayne M. Patrick. 2016. "Mechanistic and Evolutionary Insights from the Reciprocal Promiscuity of Two Pyridoxal Phosphate-Dependent Enzymes." *Journal of Biological Chemistry* 291(38): 19873–87. <http://dx.doi.org/10.1074/jbc.M116.739557>.
- Sourisseau, Marion et al. 2019. "Deep Mutational Scanning

- Comprehensively Maps How Zika Envelope Protein Mutations Affect Viral Growth and Antibody Escape” ed. Julie K. Pfeiffer. *Journal of Virology* 93(23): 1–17.
<https://journals.asm.org/doi/10.1128/JVI.01291-19>.
- Spies, M. Ashley, and Michael D. Toney. 2003. “Multiple Hydrogen Kinetic Isotope Effects for Enzymes Catalyzing Exchange with Solvent: Application to Alanine Racemase.” *Biochemistry* 42(17): 5099–5107.
- . 2007. “Intrinsic Primary and Secondary Hydrogen Kinetic Isotope Effects for Alanine Racemase from Global Analysis of Progress Curves.” *Journal of the American Chemical Society* 129(35): 10678–85.
- Stamper, C. Geoffrey F., Anthony A. Morollo, and Dagmar Ringe. 1998. “Reaction of Alanine Racemase with 1-Aminoethylphosphonic Acid Forms a Stable External Aldimine.” *Biochemistry* 37(29): 10438–45.
- Starita, Lea M. et al. 2013. “Activity-Enhancing Mutations in an E3 Ubiquitin Ligase Identified by High-Throughput Mutagenesis.” *Proceedings of the National Academy of Sciences of the United States of America* 110(14).
<https://pubmed.ncbi.nlm.nih.gov/23509263/> (June 10, 2021).
- Strych, Ulrich, and Michael J. Benedik. 2002. “Mutant Analysis Shows That Alanine Racemases from *Pseudomonas Aeruginosa* and *Escherichia Coli* Are Dimeric.” *Journal of Bacteriology* 184(15): 4321–25.
- Strych, Ulrich, Hung Chung Huang, Kurt L. Krause, and Michael J. Benedik. 2000. “Characterization of the Alanine Racemases from *Pseudomonas Aeruginosa* PAO1.” *Current Microbiology* 41(4): 290–94. <https://pubmed.ncbi.nlm.nih.gov/10977898/> (June 4, 2021).
- Sultan, Marc et al. 2008. “A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome.” *Science* 321(5891): 956–60. <https://pubmed.ncbi.nlm.nih.gov/18599741/> (June 10, 2021).
- Sun, Shaoxian, and Michael D. Toney. 1999. “Evidence for a Two-Base Mechanism Involving Tyrosine-265 from Arginine- 219 Mutants of Alanine Racemase.” *Biochemistry* 38(13): 4058–65.

Bibliography Chapter 3.3

- Sung, C. K., H. Li, J. P. Claverys, and D. A. Morrison. 2001. "An RpsL Cassette, Janus, for Gene Replacement through Negative Selection in *Streptococcus Pneumoniae*." *Applied and Environmental Microbiology* 67(3–12): 5190–96. <http://www.tigr.org>.
- Tavassoli, Ali et al. 2008. "Inhibition of HIV Budding by a Genetically Selected Cyclic Peptide Targeting the Gag–TSG101 Interaction." *ACS Chemical Biology* 3(12): 757–64. <https://pubs.acs.org/doi/10.1021/cb800193n>.
- Tawfik, Dan S. 2010. "Messy Biology and the Origins of Evolutionary Innovations." *Nature Chemical Biology* 6(10): 692–96. <http://dx.doi.org/10.1038/nchembio.441>.
- Thibessard, Annabelle et al. 2001. "Hydrogen Peroxide Effects on *Streptococcus Thermophilus* CNRZ368 Cell Viability." *Research in Microbiology* 152(6): 593–96. <https://hal.univ-lorraine.fr/hal-01631261> (June 22, 2021).
- Tinberg, Christine E. et al. 2013. "Computational Design of Ligand-Binding Proteins with High Affinity and Selectivity." *Nature* 501(7466): 212–16.
- Di Tommaso, Paolo et al. 2011. "T-Coffee: A Web Server for the Multiple Sequence Alignment of Protein and RNA Sequences Using Structural Information and Homology Extension." *Nucleic Acids Research* 39(SUPPL. 2): W13–17. <http://www.playframework.org> (June 6, 2021).
- Tóth-Petróczy, Ágnes, and Dan S. Tawfik. 2011. "Slow Protein Evolutionary Rates Are Dictated by Surface - Core Association." *Proceedings of the National Academy of Sciences of the United States of America* 108(27): 11151–56. www.pnas.org/cgi/doi/10.1073/pnas.1015994108 (June 21, 2021).
- Townend, Jaime E., and Ali Tavassoli. 2016. "Traceless Production of Cyclic Peptide Libraries in *E. Coli*." *ACS Chemical Biology* 11(6): 1624–30.
- Traxlmayr, Michael W. et al. 2012. "Construction of a Stability Landscape of the CH3 Domain of Human IgG1 by Combining Directed Evolution with High Throughput Sequencing." *Journal of Molecular Biology* 423(3): 397–412. <https://pubmed.ncbi.nlm.nih.gov/22846908/>

(June 10, 2021).

- Wagenaar, Timothy R. et al. 2014. "Resistance to Vemurafenib Resulting from a Novel Mutation in the BRAFV600E Kinase Domain." *Pigment Cell and Melanoma Research* 27(1): 124–33.
- Walsh, C. T. 1989. "Enzymes in the D-Alanine Branch of Bacterial Cell Wall Peptidoglycan Assembly." *Journal of Biological Chemistry* 264(5): 2393–96.
- Wang, Sheng, Wei Li, Shiwang Liu, and Jinbo Xu. 2016. "RaptorX-Property: A Web Server for Protein Structure Property Prediction." *Nucleic Acids Research* 44.
<https://academic.oup.com/nar/article/44/W1/W430/2499323>.
- Watababe, Akira et al. 1999. "Role of Lysine 39 of Alanine Racemase from *Bacillus Stearothermophilus* That Binds Pyridoxal 5'-Phosphate: Chemical Rescue Studies of Lys39 → Ala Mutant." *Journal of Biological Chemistry* 274(7): 4189–94.
- Watanabe, Akira et al. 2002. "Reaction Mechanism of Alanine Racemase from *Bacillus Stearothermophilus*: X-Ray Crystallographic Studies of the Enzyme Bound with N-(5'-Phosphopyridoxyl)Alanine." *Journal of Biological Chemistry* 277(21): 19166–72.
- Watanabe, Akira, Yoichi Kurokawa, Tohru Yoshimura, and Nobuyoshi Esaki. 1999. "Role of Tyrosine 265 of Alanine Racemase from *Bacillus Stearothermophilus*." *Journal of Biochemistry* 125(6): 987–90.
- Watanabe, Akira, Tohru Yoshimura, Bunzo Mikami, and Nobuyoshi Esaki. 1999. "Tyrosine 265 of Alanine Racemase Serves as a Base Abstracting α -Hydrogen from L-Alanine: The Counterpart Residue to Lysine 39 Specific to D-Alanine." *Journal of Biochemistry* 126(4): 781–86.
- Waterhouse, Andrew et al. 2018. "SWISS-MODEL: Homology Modelling of Protein Structures and Complexes." *Nucleic Acids Research* 46(W1): W296–303.
- Wei, Qiong et al. 2010. "Testing Computational Prediction of Missense Mutation Phenotypes: Functional Characterization of 204 Mutations of Human Cystathionine Beta Synthase." *Proteins* 78(9): 2058. /pmc/articles/PMC3040297/ (August 12, 2021).

Bibliography Chapter 3.3

- Wells, James A. 1990. "Additivity of Mutational Effects in Proteins." *Biochemistry* 29(37): 8509–17.
- Whitehead, Timothy A. et al. 2012. "Optimization of Affinity, Specificity and Function of Designed Influenza Inhibitors Using Deep Sequencing." *Nature Biotechnology* 30(6): 543–48.
<http://dx.doi.org/10.1038/nbt.2214>.
- Wu, Dalei et al. 2008. "Residues Asp164 and Glu165 at the Substrate Entryway Function Potently in Substrate Orientation of Alanine Racemase from E. Coli : Enzymatic Characterization with Crystal Structure Analysis ." *Protein Science* 17(6): 1066–76.
- Wu, N. C. et al. 2013. "Systematic Identification of H274Y Compensatory Mutations in Influenza A Virus Neuraminidase by High-Throughput Screening." *Journal of Virology* 87(2): 1193–99.
- Xie, Zhoujie, Fengxia Qi, and Justin Merritt. 2013. "Development of a Tunable Wide-Range Gene Induction System Useful for the Study of Streptococcal Toxin-Antitoxin Systems." *Applied and Environmental Microbiology* 79(20): 6375–84.
- Xue, Zhangwei et al. 2013. "Characterization and Preliminary Mutation Analysis of a Thermostable Alanine Racemase from Thermoanaerobacter Tengcongensis MB4." *Extremophiles* 17(4): 611–21.
- Yang, Jian Rong, Ben Yang Liao, Shi Mei Zhuang, and Jianzhi Zhang. 2012. "Protein Misinteraction Avoidance Causes Highly Expressed Proteins to Evolve Slowly." *Proceedings of the National Academy of Sciences of the United States of America* 109(14): 5158.
<https://www.pnas.org/content/109/14/E831> (June 21, 2021).
- Yeh, So Wei et al. 2014. "Local Packing Density Is the Main Structural Determinant of the Rate of Protein Sequence Evolution at Site Level." *BioMed Research International* 2014.
- Yokoigawa, Kumio et al. 2001. "Gene Cloning and Characterization of Alanine Racemases from Shigella Dysenteriae, Shigella Boydii, Shigella Flexneri, and Shigella Sonnei." *Biochemical and Biophysical Research Communications* 288(3): 676–84.
- Young, Travis S. et al. 2011. "Evolution of Cyclic Peptide Protease Inhibitors." *Proceedings of the National Academy of Sciences of the*

United States of America 108(27): 11052–56.

- Yow, Geok Yong, Akira Watanabe, Tohru Yoshimura, and Nobuyoshi Esaki. 2003. "Conversion of the Catalytic Specificity of Alanine Racemase to a D-Amino Acid Aminotransferase Activity by a Double Active-Site Mutation." *Journal of Molecular Catalysis B: Enzymatic* 23(2–6): 311–19.
- Zettler, Joachim, Vivien Schütz, and Henning D. Mootz. 2009. "The Naturally Split Npu DnaE Intein Exhibits an Extraordinarily High Rate in the Protein Trans-Splicing Reaction." *FEBS Letters* 583(5): 909–14. <https://pubmed.ncbi.nlm.nih.gov/19302791/> (June 19, 2021).
- Zhang, Jianzhi, and Jian Rong Yang. 2015. "Determinants of the Rate of Protein Sequence Evolution." *Nature Reviews Genetics* 16(7): 409–20. <http://dx.doi.org/10.1038/nrg3950>.
- Zhang, Linlin et al. 2020. "Crystal Structure of SARS-CoV-2 Main Protease Provides a Basis for Design of Improved α -Ketoamide Inhibitors." *Science* (PG-): eabb3405. NS -.
- Zhou, Hang et al. 2012. "Xylose Isomerase Overexpression along with Engineering of the Pentose Phosphate Pathway and Evolutionary Engineering Enable Rapid Xylose Utilization and Ethanol Production by *Saccharomyces Cerevisiae*." *Metabolic Engineering* 14(6): 611–22.